Università degli Studi di
Modena e Reggio Emilia

Multiscale Modeling, Computational Simulations and Characterization in Material and Life Sciences
PhD cycle XXVI

A computational approach for the identification of genome cis-regulatory elements from ChIP-seq data

PhD student:
Guidantonio Malagoli Tagliazucchi

Tutor:
Prof. Mauro Leoncini
Co-tutor:
Prof. Silvio Bicciato
Coordinatore del Dottorato:
Prof. Daniele Malferrari
Direttore della Scuola:
Prof. Ledi Menabue

Abstract

The genome-wide identification of cis-regulatory elements (CRE), i.e., those DNA sequences, as promoters, enhancers, insulators, and silencers, required to regulate gene expression, still represents a major challenge for computational biology. Several types of cis-regulatory elements are present in a genome, all with different characteristics and functions. In particular, promoters are DNA sequences that, first, direct RNA polymerase to initiate at the transcription start site (TSS) and then promote transcription. Enhancers are DNA regions that increase the expression level of their target gene (or genes). Differently from promoters, enhancers drive gene expression at distance and are defined as tissue-specific cis-regulatory elements. Distinct, specialized enhancers, uncoupled from promoters and positioned away from the transcription start site, coordinate the transcription of a given gene in multiple tissues, in response to distinct signaling cascades, and at different stages during differentiation.

Promoters and enhancers are co-localized with different proteins, such as transcription factors (TF) or histone modifications (HM), whose identification is required to characterize cis-regulatory modules. However, while it is well accepted that active-promoters co-localize in regions with low levels of specific histone modifications, histone-marks or proteins that co-localize with enhancers remain only partially understood. Recently, Chromatin immuno-precipitation (ChIP), followed by high-throughput DNA sequencing (ChIP-seq) and bioinformatics analyses, has become a valuable and widely-used approach to map the genomic location of transcription-factor binding and histone modifications in living cells, i.e., to identify cis-regulatory modules. Since its introduction, several bioinformatics approaches have been developed to identify genome cis-regulatory elements from ChIP-seq data, as Hidden Markov Models, dynamic Bayesian networks, and profile methods. All these approaches require ChIP-seq data for a variety of histone marks, but a consensus on the optimal combination of histone modifications and algorithmic solutions for an efficient identification of CRE has not been reached, yet.

Here, we present a computational procedure based on support vector machine (SVM) models and ChIP-seq data to i) determine the optimal number of histone marks necessary for the genome-wide identification of cis-regulatory elements, and ii) detect promoters and enhancers through the integration of ChIP-seq data with Cap Analysis of Gene Expression profiles. The bioinformatics procedure has been designed, developed, and tested using ChIP-seq data obtained from various cell types (hematopoietic, neuronal, endothelial, epithelial) at different differentiation stages (stem cells, progenitors, terminally differentiated cells) and provided a unique collection of genes and regulatory regions involved in self-renewal, commitment, and differentiation of human stem cells and their progeny in different tissues. The procedure has been coded in an R package (*SVM2CRM*) and the computational performances have been optimized using libraries specifically designed for the analysis of large data

Sommario

I genomi contengono differenti sequenze regolatrici, quali promotori, enhancers, insulators e silencers, che sono cruciali nel controllare il livello di espressione dei geni. In particolare, i promotori sono sequenze di DNA che dirigono l'RNA polimerasi ad iniziare la trascrizione a livello del sito di inizio della trascrizione; la loro funzione è quindi di promuovere la trascrizione di un gene (o di più geni). Gli enhancers sono invece regioni di DNA che permettono di aumentare l'espressione genica di un loro gene target (o di più geni target). Gli enhancers si trovano lontani dal sito d'inizio della trascrizione e, diversamente dai promotori, sono definiti come elementi cis-regolatori tessuto-specifici. Enhancers specializzati, indipendentemente dal promotore, coordinano la trascrizione di un gene in tessuti diversi in risposta a differenti cascate di segnali cellulari e in differenti stadi del differenziamento cellulare.
I promotori e gli enhancers sono caratterizzati dalla presenza di fattori trascrizionali e da specifiche modificazioni istoniche, la cui identificazione è cruciale per definire in maniera completa i moduli regolativi. Tuttavia, mentre è noto che i promotori attivi risiedono frequentemente in regioni che presentano bassi livelli di alcune modificazioni istoniche, non è ancora chiaro quali siano i marcatori istonici o le proteine che possono permettere l'identificazione e la localizzazione degli enhancers. Recentemente, l'immunoprecipitazione della cromatina seguita dal sequenziamento massivo del DNA immunoprecipitato (ChIP-seq) è diventata la tecnica di eccellenza per localizzare i siti di legame dei fattori trascrizionali, le modificazioni istoniche e la localizzazione dei moduli cis-regolatori. Per l'analisi dei dati di ChIP-seq e l'identificazione degli elementi cis-regolatori sono stati sviluppati diversi algoritmi e metodi bioinformatici, quali ad esempio modelli nascosti di Markov, metodi basati sulla statistica Bayesiana e tecniche di *pattern matching*. Benché tutti questi algoritmi richiedano dati di ChIP-seq per diverse modificazioni istoniche, non è ancora stato trovato un consenso sulla combinazione ottimale di modificazioni istoniche e soluzioni algoritmiche che consenta un'efficiente identificazione dei moduli di regolazione trascrizionale.
In questo lavoro viene presentata una procedura computazionale basata su Support Vector Machine per l'analisi di dati di ChIP-seq al fine di i) determinare il numero ottimale di modificazioni istoniche necessario per l'identificazione in tutto il genoma degli elementi cis-regolatori e ii) identificare promotori ed enhancers mediante l'integrazione di dati di ChIP-seq con profili genici ottenuti da Cap Analysis of Gene-Expression. Le analisi bioinformatiche sono state progettate e sviluppate utilizzando dati di ChIP-seq ottenuti da vari tipi cellulari (cellule ematopoietiche, neuronali, endoteliali, epiteliali) a differenti stadi di differenziamento (staminali, progenitori, cellule terminalmente differenziate) e hanno fornito come risultato una collezione unica di geni e regioni regolatorie coinvolti nei processi differenziativi di cellule staminali umane e della loro progenie terminalmente differenziata. Gli algoritmi e le funzioni utilizzate in questo lavoro sono stati assemblati in un pacchetto R, denominato *SVM2CRM*, le cui performance computazionali sono state ottimizzate utilizzando specifiche librerie per l'analisi di dati di grandi dimensioni.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In 2001 the *Human Genome Sequencing Consortium* released the first draft of the complete human genome sequence. The availability of the complete human genome allowed the researchers to further understand the underlying molecular mechanisms of diseases and thereby facilitated the rational design of new drugs and diagnostics tools targeted at those mechanisms (McPherson et al., 2001). However, despite intensive studies, especially in identifying protein-coding genes, our understanding of the genome is far from complete, particularly with regard to non-coding RNAs, alternatively spliced transcripts, and regulatory sequences. The complete analysis of these regulatory sequences, and of the mechanisms that regulate transcription, is essential and is a mandatory requirement to fully understand genome organization, and ultimately cell biology, in physiological and pathological states.

To fully address these issues, recently the scientific community working on genomes created the ENCyclopedia Of DNA Elements (ENCODE). The main objective of this project is to catalog structural and functional components encoded in the genome sequences of higher organisms as mouse or humans. In this effort, the ENCODE project used next generation sequencing technologies, and specifically, techniques like RNA-sequencing and Cap Analysis of Gene expression (CAGE) to quantify gene expression and discovery new gene isoforms, and chromatin immuno-precipitation sequencing (ChIP-seq) to identify genomic functional regions, i.e., DNA stretches that bind proteins and cooperate in the genome utilization. ENCODE also developed several methods to analyze and integrates sequence-based data (ChIP-seq, CAGE).

The widely usage of high throughput sequencing technologies and the boost impressed by the ENCODE project led the researchers in epigenetics and epigenomics to identify *functional sequences* in different cellular context. This is an important challenge of biology because the comprehensively and accurately identification of the DNA sequences that are required to regulate gene expression, i.e. *functional sequences* or *cis-regulatory elements*, can

elucidate the underlying mechanisms that regulate cell differentiation into distinct tissues and organs. In the genome, several kinds of *cis-regulatory elements* are present, as promoters, enhancers, insulators, silencers, all with different characteristics and functions. Promoters are DNA sequences that direct RNA polymerase to initiate at the transcription start site (TSS) and then promote the transcription. Enhancers are DNA regions that increase gene expression of their target gene (or genes). In particular, enhancers have the characteristic to drive gene expression at distance. Their uncoupling from promoters, and the positioning away from the transcription start site, permits a given gene to be transcribed in multiple tissues, at varied levels in response to distinct signaling cascades, and at different stages during development via the utilization of distinct, specialized enhancers. Differently from promoters, enhancers are defined as tissue-specific cis-regulatory elements. Silencers have an opposite effect on a target gene, as compared to enhancers; indeed silencers reduce gene expression levels. Finally, insulators are DNA sequences that control the ability of an enhancer to regulate a promoter by an enhancer blocking activity. All these cis-regulatory elements are co-localized with different proteins, such as transcription factors (TF) or histone modifications (HM), and the identification of these proteins is mandatory for the genome-wide detection of cis-regulatory modules.

The genome wide identification of the *functional sequences* that binds the proteins is achievable using chromatin immuno-precipitation followed by sequencing (ChIP-seq) and Cap Analysis of Gene Expression. In ChIP-seq, specific regions of cross-linked chromatin, which is genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then subjected to high throughput sequencing to determine the regions in the genome most often bound by the protein to which the antibody was direct. As such, ChIP-seq allows accessing fundamental components of the transcription control, as for instance the genome-wide localization of DNA methylation and covalent modifications of histone proteins, also known as epigenetic modifications. The identification of promoters and enhancers (in particular those that present the transcription of a new class of non-coding RNA (eRNA)) can be addressed using Cap Analysis of Gene Expression. CAGE implies the capture of the methylated 5' end of RNA formed at the initiation of transcription, followed by its high-throughput sequencing. Differently from microarray, CAGE identifies not only the expression levels of an mRNA, but also the localization of each transcription start site (TSSs, Shiraki et al., 2003). This allows the reconstruction of a genome-wide network of transcriptional regulation, by searching the promoters regions surrounding each CAGE peak and defining transcription start site for potential transcription factor binding sites. In summary, CAGE and ChIP-seq allow characterizing the entire cell transcriptome and epigenome, respectively. Any of the above mentioned techniques produces data that are highly informative *per se*, but merging and integrating the various types of information would offer the potential to answer many long standing questions related to fundamental mechanisms of gene and genome regulation. The understanding of transcriptional regulation in higher eukaryotes is severely hampered by the limited number of known regulatory elements, particularly long-range regulatory modules, such as enhancers and enhancer-blocking elements (insulators). Although only few regulatory elements have been identified, nevertheless their genomic patterns are well characterized and can, in principle, represent a fingerprint to identify novel elements. As an example, known enhancers have a specifics chromatin signatures that can be used to detect new enhancers or new classes of non-coding elements.

1.2 Contribution

With the increase number of epigenetic studies, bioinformatics is facing a number of challenges related to an overwhelming mass of genomic data that need to be managed, integrated, analyzed, and exploited. In particular, the integration of different sources of information (e.g., ChIP-seq and CAGE data) and the identification of cis-regulatory elements still present unsolved computational issues.

At the state of art, data integration can be addressed in three major ways, i.e., *reduction of data complexity*, *unsupervised integration, and supervised integration. Reduction of data complexity* is a mandatory step when are considering technologies, as ChIP-seq, that produce, in any single experiment, millions of short sequence reads, mapping the entire genome to a continuous signal. A simple approach to reduce data dimensionality from millions of points to more manageable hundreds or thousands of sites is to summarize each experiment as a collection of genomic regions with strong signal enrichment. Specifically, with ChiIP-Seq data, this can be achieved first using peak-calling algorithms that discretize the genome into regions with and without signal enrichment and then intersecting enriched regions from different experiments. The *unsupervised integration* is based on unsupervised learning (i.e., clustering), in which data are queried with no prior biases, knowledge, or hypotheses. The reduction step is then performed identifying consistent patterns in different data sets. In this framework, the genome structure serves as a scaffold on which high-throughput data are assembled and clustering is used, first, to classify genomic loci into conceptual modules with shared attributes and, then, to connect persistent modules. For example, clustering of RNA expression reveals co-expressed genes while clustering of histone modifications returns loci that share similar chromatin-structures. The connection of these modules can be obtained in several ways as, for instance, examining a module derived from the analysis of one type of data (e.g., chromatin signatures) in the context of another data type (e.g., DNA methylation). Another approach for inferring regulatory modules requires concepts of network biology, which leverages high-throughput techniques to find relationships that connect genomic loci and conceptual groups, as, for example, in the quest for genes under the regulation of a given transcription factor. In this specific case, clustering of ChIP-seq peaks would identify a subgroup of putative binding sites for a given transcription factor sharing a similar chromatin signature while RNA-seq data would confirm if target genes of the transcription factor (characterized by the subgroup of putative binding sites) are highly expressed or not. This second level of integration, linking different kinds of experiments, can form a knowledge base that can be used to provide biological insights or to formulate hypotheses for further studies. Accordingly unsupervised integration is a useful discovery tool to search correlations between two or more experiments. The novel associations can lead to new hypotheses of function, which can be followed up by supervised integration. In this way, high-throughput experiments are viewed as read-outs to identify emerging, unexpected associations. Differently, *supervised integration* relies on testable hypotheses and usually starts with a prediction based on an observation and ends with a test of this prediction. Correlation analysis and Bayesian probabilistic networks are among the tools of choice for performing supervised analyses, although both methods can become computationally intensive as the size of the dataset grows (Hawkins RD et al., 2010). The three approaches described above are commonly employed in very simple bioinformatics workflows in which only a single type of data is analyzed.

When considering the genome wide identification of *cis-regulatory elements*, different techniques have been developed to identify *functional elements* and, among others, chromatin immuno-precipitation followed by short-tag sequencing (ChIP-seq) has become one of the most used. Several bioinformatics methodologies have been designed to analyze ChIP-seq

data and characterize genome regulatory elements, as Hidden Markov Models (Ernst et al., 2011), dynamic Bayesian networks (Hoffman et al., 2012), and profile methods (Hentzman et al., 2007). Despite all efforts, no standard method has been developed to detect cis-regulatory yet, mostly due to different experimental designs (numbers of analyzed histone marks) and to a lack of consensus on the rules to detect cis-regulatory modules. In fact, while it is well accepted that active-promoters co-localize in regions with low levels of H3K4me1 (mono-methylation of lysine 4 on histone 3) and high levels of H3K4me3 (tri-methylation of lysine 4 on histone 3), it is not clear which histone-marks or proteins co-localize with enhancers. The problem is further complicated by the fact that enhancers are tissue-specific. In general high levels of H3K4me1 and low levels of H3K4me3 characterize strong-enhancers, but other histone marks like H3K27ac (acetylation of lysine 27 on histone 3), H3K4me3, and H3K9ac (acetylation of lysine 9 on histone 3) may be informative to detect enhancers.

In this context, the methodological aim of this thesis was to develop a bioinformatics framework for the integration of genomics data obtained from CAGE and ChIP-seq. The applicative aim of the work was to map and classify novel regulatory elements and bi-directionally transcribed non-coding RNAs (eRNA) in the human genome through the integrative analysis of genomic data from different types of progenitors and terminally differentiated cells. The research work resulted in:

a) a pipeline to perform *reduction of data complexity integration* using ChIP-seq data;
b) an R-based pipeline to perform *reduction of data complexity integration between* ChIP-seq and CAGE data. Standard data analyses was performed for any type of experiment but the outputs were merged into a unique, integrated data structure that can be inspected to identify patterns of genomic signals and to extract biological information;
c) a custom R-workflow, named *intersect-ratio-count-method* (IRCM), to identify promoter and enhancers using specific histone marks (i.e. H3K4me1 and H3K4me3) and to perform the integrative analysis of ChIP-seq and CAGE data;
d) an R-package, named *SVM2CRM* and based on machine-learning methods, to find the optimal number of histone marks for an efficient prediction of promoters and enhancers.

1.3 Document organization

Chapter 2 contains a description of material and methods used in this thesis. In particular, Section 2.1 provides an extensive description about high order chromatin organization and cis-regulatory elements as promoters and enhancers; Sections 2.2 and 2.3 provide a deeper explanation about the chromatin immune precipitation sequencing and Cap-Analysis of Gene Expression; Sections 2.4 and 2.6 describe the intersect reads count and the SVM2CRM methods for the detection of cis-regulatory elements. Chapter 3 summarizes all results obtained from the analysis of the various genomic data in the various cell types using the bioinformatics approaches presented in Chapter 2. Finally, conclusions and future perspectives of this work are drawn in Chapter 4.

Chapter 2

Materials and Methods

This chapter contains a first section describing the hierarchical organization of chromatin and the functional elements that regulate the expression of promoters and enhancers during cell differentiation. The subsequent section describes methodological aspects of chromatin immuno-precipitation sequencing and Cap Analysis of Gene Expression (CAGE). These technologies allow performing genome-wide mapping of cis-regulatory elements and studying the regulation of gene expression inside a cell. Moreover, details on the theoretical and applicative aspects of the bioinformatics analysis of ChIP-seq and CAGE data are provided. This chapter also describes the computational methodologies used in this work for the integration of ChIP-seq and CAGE data, the annotations of CAGE peaks, and the predictions of *cis-regulatory elements* using the various computational pipelines (i.e., intersect-ratio-count-method (IRCM) and the machine learning method, based on support vectors machines (SVM2CRM)). The final section describes the ChIP-seq and CAGE datasets used throughout this work.

2.1 Higher-order chromatin organization

Eukaryotic chromatin structure can be viewed as a series of imposed organizational layers, where, at the root, there is the DNA sequence and its direct chemical modification by cytosine methylation. DNA is folded around nucleosomes and chromatin, the DNA-nucleosome polymer, is a dynamic molecule that exists in many configurations (Figure 1). Chromatin can be classified as either euchromatic or heterochromatic. Euchromatin is the de-condensed form of chromatin and may be transcriptionally active or inactive; heterochromatin is instead the highly compacted and silenced form of chromatin. Heterochromatin may exist as permanently silent chromatin (constitutive heterochromatin), and in this case genes will rarely be expressed by any cell type, or repressed (facultative heterochromatin), containg genes that may be transcribed in some cells during a specific cell cycle or developmental stage.

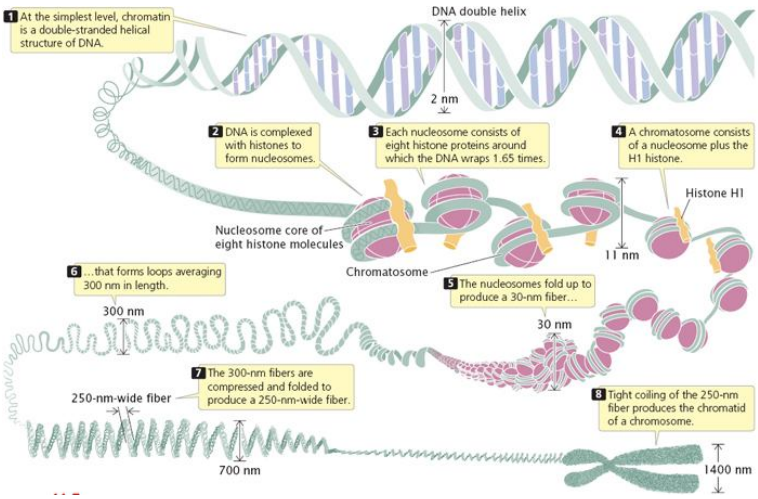

Figure 1. High-order chromatin organization

Thus, chromatin is a dynamic macromolecular structure that may assume a spectrum of states. Nucleosomes are arranged in an 11-nm template, a structure that represents an active and largely *unfolded* interphase configuration with DNA periodically wrapped around repeating nucleosome units. The 11-nm template can be altered by *cis*-effects and *trans*-effects of covalently modified histone tails. A well-known example of *cis*-effect is the histone acetylation that, neutralizing positive charges of highly basic histone tails, generates a localized expansion of the chromatin fiber thus enabling a better access of the transcription machinery to the double helix. Configuration changes by *trans*-effects are mediated by the recruitment of modification-binding partners to the chromatin. Modification-binding partners act as *readers* of a specific histone mark. A second higher order chromatin structure is the 30-nm fiber, i.e., a compact and repressive state that can be achieved through the recruitment of linker histone H1 and/or modification-dependent architectural chromatin-associated factors as heterochromatin 1 (HP1) or polycomb (PC). The final organization into looped chromatin domains (300-700 nm) is obtained anchoring the chromatin fiber to the nuclear periphery or other nuclear lamins.

2.1.1 Chromatin and chromatin modifications

Chromatin is the state in which DNA is packed within a cell. The nucleosome is the fundamental unit of chromatin and it is composed by an octamer with four cores (H3, H4, H2A, H2B) around which 147 base pairs of DNA are wrapped. The core histone proteins that make up the nucleosome are small and highly basic. They are composed of a globular domain and flexible (relatively unstructured) histone tails (N-terminal) which protrude from the surface of the nucleosome. Based on amino acid sequence, histone proteins are highly conserved from yeast to humans. A striking feature of histones, and particularly of their tiles, is the large number and type of modified residues they possess. There are least eight distinct types of modifications that can be found on histones, among which the most studied are small covalent modifications as acetylation, methylation, and phosphorylation (Table 1). Histones are modified at many sites: there are more than 60 different residues on histone where modifications have been detected either by specific antibodies or by mass spectrometry. However, this represents a huge underestimate of the number of modifications that can take place on histones. Indeed, an extra complexity is added by the fact that lysine or arginine methylation may occur in the form of mono-, di-, or tri-methylation.

Table 1. List of chromatin modifications

Chromatin Modifications	Modified residues	Regulated function
Acetylation	**K**-ac	Transcription, Repair, Replication, Condensation
Methylation (lysines)	**K**-me1**, K**-me2**, K**-me3	Transcription, Repair
Methylation (arginines)	**R**-me1**, R**-me3a**, R**-me2s	Transcription
Phosphorylation	**S**-ph**, T**-ph	Transcription, Repair, Condensation
Ubiquitylation	**K**-ub	Transcription, Repair
Sumoylation	**K**-su	Transcription
ADP ribosylation	**E**-ar	Transcription
Deimination	**R > Cit**	Transcription
Proline isomerization	**P-cis > P-trans**	Transcription

This vast array of modification gives enormous potential for functional responses, but it has to be remembered that not all these modifications will be on the same histone at the same time. The appearance of a modification will depend on the signaling conditions within the cells and thus histone modifications are dynamic and rapidly changing. Acetylation, methylation, phosphorylation, and deimination can appear and disappear on chromatin within minutes after a stimulus is arrived at the cell surface. As such, examining bulk histones under on specific set of conditions (with either antibodies or mass spectrometry) will identify only a portion of all possible modifications. Finally, it is commonly assumed that each individual modification leads to a biological consequence. However, proof of a consequence is not always easy to provide and is often based on correlation: a modification appears on a gene under certain conditions (e.g., when it is transcribed) and disappears when its state is reversed (e.g., when the gene is silenced). Proving causality for a modification involves showing that the catalytic activity of the enzyme that mediates the modification is necessary for the biological response. Furthermore, there may be signaling redundancy such that more than one enzyme may be capable of modifying a specific site. In this case, the effects of inactivating one enzyme may be masked by an up regulation in the activity of a second distinct, but related, enzyme.

2.1.2 Transcriptional regulatory elements

The expression of eukaryotic protein-coding genes can be regulated at several steps, including transcription initiation and elongation, and mRNA processing, transport, translation, and stability. Most regulation, however, is believed to occur at the level of transcription initiation. In eukaryotes, transcription of protein-coding genes is performed by RNA polymerase II. Genes transcribed by RNA polymerase II typically contain two distinct families of cis-acting transcriptional regulatory DNA elements: i) promoters, which is composed of a core promoter and nearby (proximal) regulatory elements, b) distal regulatory elements which can be enhancers, silencers, insulators, or locus control regions (LCR) (Figure. 2). This *cis*-acting regulatory sequences contain recognition sites for *trans*-acting DNA-binding transcription factors, which either enhance or repress transcription.

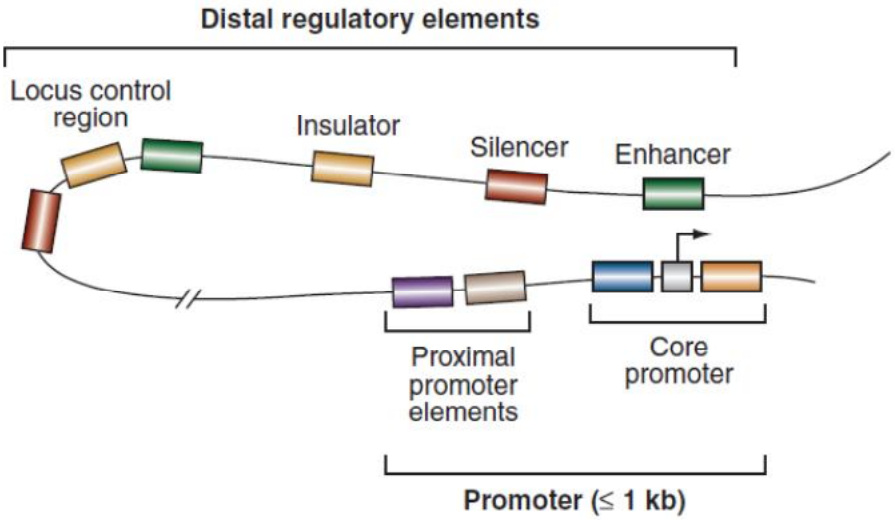

Figure 2. Schematic of a typical gene regulatory region. The promoter, which is composed of a core promoter and proximal promoter elements, typically spans less than 1 kb pairs. Distal (in this case upstream) regulatory elements, which can include enhancers, silencers, insulators and locus control regions, can be located up to 1 Mb pairs from the promoter. These distal elements may contact the core promoter or proximal promoter through a mechanism that involves looping out the intervening DNA

2.1.3 Proximal regulatory elements

Core and proximal promoters are the basic elements that drive transcription and they may be sufficient for transcription of those genes that are expressed ubiquitously and involved in housekeeping functions. The core promoter is the region at the start of a gene that serves as anchor point for the recruitment of the general transcriptional machinery and the pre-initiation complex (PIC) assembly, and defines the position of the transcription start site (TSS) as well as the direction of transcription. The proximal promoter is defined as the region immediately upstream (up to a few hundred base pairs) from the core promoter, and typically contains multiple transcription factor binding sites (TFBS, short degenerate sequences of 6-20 base pairs).

2.1.4 Distal regulatory elements

Temporal and tissue-specific gene expression in mammals depends primarily on distal regulatory elements (i.e., enhancers, silencers, insulators, and LCR), which are often located far away from the genes they control. Enhancers were identified as regions of the genome that could markedly increase the transcription of a gene. They usually regulate transcription in a spatial or temporal specific manner, and they function independent of both the distance from and orientation relative to the promoter; enhancer may in fact reside several hundred kilo base pairs (kb) upstream of a promoter, downstream of a promoter in an intron, or even beyond the 3' end of the gene. Enhancers are typically composed of a cluster of TFBSs: enhancer-bound transcription activators act looping out the intervening sequences and contacting the promoter region to facilitate the assembly of the transcription machinery at gene promoters, and they can also recruit histone modifying enzymes or ATP-dependent chromatin remodeling complexes to alter chromatin structure and increase the accessibility of the DNA to other proteins. In addition, it appears that enhancer sequences themselves are not mere binding sites for transcription factors but are also transcribed into non-coding RNAs, which may have an active role in the control of transcription by stabilizing long-range enhancer-promoter interactions.

Silencers, on the other hand, are sequence-specific elements that confer a negative (i.e., silencing or repressing) effect on the transcription of a target gene. They function independently of orientation and distance from the promoter and they can be located as part of a proximal promoter, as part of a distal enhancer, or as an independent distal regulatory module; in this regard, silencers can be located far from their target gene, in its intron, or in its 3'-untranslated region. Silencers bind transcription repressors acting through inhibition of gene transcription; a number of models have been proposed for repressor function. In some cases, repressors appear to function by blocking the binding of a nearby activator, or by directly competing for the same site. Alternatively, a repressor may prevent the general transcriptional machinery from accessing a promoter by establishing a repressive chromatin structure through the recruitment of histone-modifying activities. Finally, it was recently suggested that a repressor might block transcription activation by inhibiting PIC assembly. Other regulatory DNA elements that contribute to the formation and maintenance of active or inactive transcription programs and play an integral part in gene regulation are insulators (also known as boundary elements). They function in a position-dependent, orientation-independent manner to block genes from being affected by the transcriptional activity of neighboring genes. They thus create boundaries in chromatin limiting the action of transcriptional regulatory elements into defined domains, and partition the genome into discrete realms of expression. There are two types of insulators: enhancer-blocking insulators, which prevent communication between discrete sequence elements (typically enhancers, or even silencer, and promoters) when positioned between them, and barrier insulators, which prevent the spread of heterochromatin. The core insulator fragment contains binding sites for several transcription factors; in vertebrates, the only known insulator protein is CCCTC-Binding Factor (CTCF) that is necessary for enhancer-blocking activity. It is proposed that CTCF creates distinct loop domains, in which the enhancer in one loop is unable to contact a promoter in a different loop. Finally, developmental and cell lineage-specific regulation of gene expression relies upon Locus control regions (LCRs), that are groups of regulatory elements involved in regulating an entire locus or gene cluster. Locus control regions are operationally defined as elements that enhance the expression of linked genes to physiological levels in a tissue-specific, position-independent and copy-number-dependent manner. LCRs are often marked by a cluster of nearby DNase I hypersensitive sites (HS) and are thought to provide an open-chromatin domain for genes to which they are linked. LCRs are typically composed of multiple *cis*-acting elements, including enhancers, silencers, insulators, and nuclear-matrix or chromosome scaffold attachment regions (MARs or SARs). These elements commonly co-localize to sites of DNAse I hypersensitivity and are bound by transcription factors (both tissue-specific and ubiquitous), co-activators, repressors, and/or chromatin modifiers. Each of the components differentially affects gene expression, and it is their collective activity that functionally defines a LCR and confers proper spatial/temporal gene expression. The final most prominent effect of the LCRs is a strong, transcription-enhancing activity. The identification of a large number of LCRs has revealed that, although LCRs are typically located upstream of their target gene(s), they can also be found within an intron of the gene they regulate, downstream of the gene, or even in the intron of a neighboring gene.

2.1.5 Chromatin cis-regulatory elements

In ChIP-seq, next-generation technology is used to deep sequence the immune-precipitated DNA molecules and thereby produce digital maps of ChIP enrichment. Zhao and coworkers, for examples, used ChIP-seq to investigate genome-wide 39 different histone

methylation and acetylation marks in human CD4+ T cells. These maps have been associated to particular modifications with various genomic features, including promoters, enhancers, and insulators and have been associated to the regulation of gene-expression. In particular, these maps highlighted that:

- most promoters coincide with regions of high GC content and CpG ratios, or CpG islands. These have been termed high CpG-content promoters (HCPs), in contrast to low CpG-content promoters (LCPs). HCPs and LCPs have different histone modification patterns. Studies of chromatin in embryonic stem cells (ES) revealed broad targeting of histone H3 lysine 4 trimethylation (H3K4me3) to virtually all HCPs, regardless of the expression state. Region of H3K4me3 were shown to be accompanied by others features of accessible chromatin, including histone acetylation and occupancy by the H3.3 histone. Active HCPs are enriched for H3K4me3 and subjected to RNA polymerase II (RNAPII) initiation. Poised HCPs are marked by the bivalent combination of H3K4me3 and H3K27me3 (Figure 3);

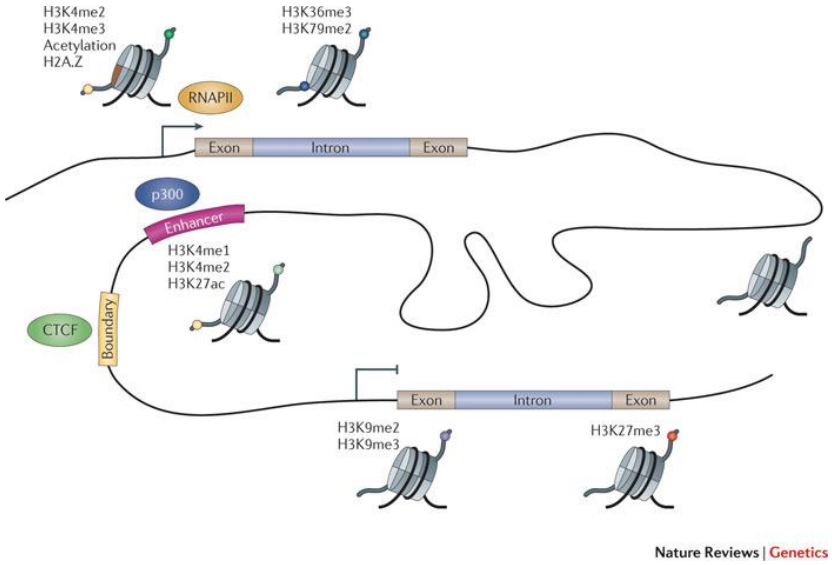

Figure 3. Specifics histone marks demarcate functional elements in mammalian genomes

- poised and repressed promoters show unique patterns of chromatin modifications that seem to reflect distinct mode of transcriptional silencing. These include H3 lysine 27 trimethylation (H3K27me3), the typical mark of polycomb repressor, H3 lysine 9 trimethylation (H3K9me3), which correlates with constitutive heterochromatin, and DNA methylation;
- mammalian genes are characterized by large numbers of exons in an expanse of introns. In many cases, alternative splicing provides an additional layer of complexity and regulation. At the DNA level, specific histone marks can distinguish primary transcripts and exons and may even have a role in determining splicing patters. This histone modifications include mono, di, and tri-methylation of various lysine residues in histone H3 (H3K36me3, H3K79me2, H4K20me1);
- enhancers are DNA elements that recruit transcription factors, RNAPII and chromatin regulators to positively influence transcription at distal promoters. Histone modification profiles have proven to be particularly useful to identifying enhancer elements in an unbiased fashion. In addition to specific histone modifications, sequence specific DNA-binding proteins and co-activators, such as p300, preferentially occupy enhancers. H3K4me1 was the first histone modification globally linked to distal regulatory elements

through genomic studies (Heintzman et al., 2007). Enhancers and promoters can be distinguished by the methylation status at the histone H3 lysine 4 whereas high levels of H3K4me3 predominantly mark promoters. These differences can be largely explained by differences in DNA sequence, with high CpG island density observed at most promoters, but not at enhancers. However, it is important to note, that the presence of H3K4me1 is not unique to enhancers and other studies suggest that enhancers also share enrichment with others histone marks. For example, the presence of H3K4me1 often precedes substantial nucleosomal depletion, H3K27ac, and enhancer activation. This signature indicates the accessibility of the genome or the chromatin dynamics at these sites. It might also reflect the physical proximity of enhancer elements to activating chromatin machinery at their target promoter through looping interactions. Although a fairly limited set of modifications, H3K4me1/3, H3K27ac, H3K27me3, is ubiquitously utilized to distinguish active and poised enhancers and promoters, these marks represent only a fragment of the full repertoire of histone modifications decorating distal regulatory regions. For example various histone marks, as H3K9ac, H3K18ac, H2BK5me1, H3K4me2, H3K9me1, H3K27me1, H3K36me1 were detected at putative enhancers. The chromatin patterns at enhancers therefore are more variable and cell type specific than chromatin patterns at promoters and insulators. In their active state, enhancers also bound by general transcriptional factors (GTFs) and RNA polymerase II (Pol II). In a recent work, Kim et al. (2010) found that RNAPII interacts with many active enhancers and transcribes bidirectional short <2 kb non coding RNAs, termed enhancer RNA (eRNAs). eRNAs can be either short, bidirectional, and non-polyadenylated or long, unidirectional, and polyadenylated, but mechanisms that specific directionality and function of eRNAs await further investigation (Natoli and Andrau, 2012). Although the function of these eRNAs remain not completely understood, nevertheless their expression levels seem to correlate with the proximal gene activity by the activation of proximal promoters;

- finally, insulators are DNA elements that block enhancer activities. They are likely related to boundary elements, which are defined by their capacity to prevent heterochromatin spreading. In mammals, the transcriptional repressor CTCF has been implicated in blocking of enhancer activity and heterochromatin spreading, and in inter-chromosomal and intra-chromosomal organization.

2.2 Chromatin immuno-precipitation sequencing

DNA-binding proteins play crucial roles in many major cellular processes, such as transcription, splicing, replication, and DNA repair. These proteins include transcriptional factors that bind preferentially to certain DNA sequences, as well as histone proteins that form the core of nucleosomes, which are the basic chromatin units. The genomic localization of either bound factors or modified histones can be accurately predicted in a particular cell type using DNA sequence alone. Chromatin immuno-precipitation coupled with microarray (ChIP-chip) or short-tag sequencing (ChIP-seq) has become the standard technique for identifying, genome-wide, the locations and biochemical modifications of bound proteins. The basic steps of the ChIP-seq assay are depicted in Figure 4. The Encyclopedia of DNA elements (ENCODE) consortium has carried out hundreds of ChIP-seq experiments and used this technique to determine a set of working standards and guidelines.

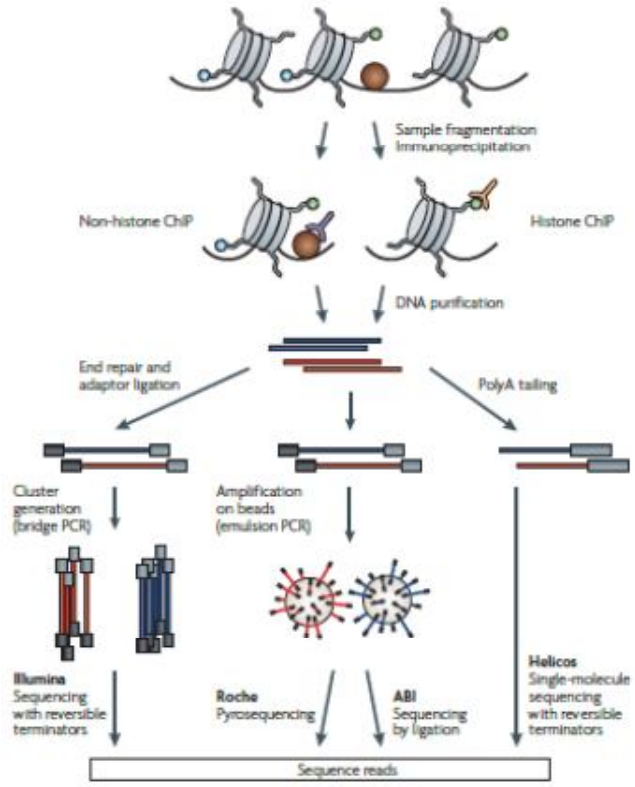

Figure 4. ChIP-seq overview

In a ChIP experiment, DNA fragments associated with a specific protein are first enriched. The DNA-binding protein is cross-linked to DNA in-vivo by treating cells with formaldehyde and the chromatin is sheared by sonication into small fragments, which are generally in the 200-600 bp range. An antibody specific to the protein of interest is used to immune-precipitate the DNA-protein complex. Finally, the crosslinks are reversed and the released DNA is assayed to determine the sequences bound by the protein. During the construction of a sequencing library, the immune-precipitated DNA is subject to size selection (typically in the ~150-300 bp range, although there seems to be a bias towards shorter fragments in sequencing).
ChIP-seq presents several advantages over other technologies to investigate DNA-binding proteins, as its enormous base pair resolution and its dynamic range. However, as all profiling technologies, also ChIP-seq can produce unwanted artifacts. Although sequencing has been reduced substantially as the technology has improved, artifacts are still present, especially towards the end of each read. This problem can be ameliorated improving alignment algorithms and computational analysis. ChIP-seq also presents a bias in fragment selection towards GC-rich content, both in library preparation and in amplification before and during sequencing. In addition, when an insufficient number of reads is generated, there is a loss of sensitivity or specificity in detection of enriched regions.

2.2.1 The analysis of ChIP-seq data

Effective analysis of ChIP-seq data requires sufficient coverage by sequence reads (sequencing depth). The required depth depends mainly on the size of the genome and the number and size of the binding sites of the proteins. For mammalian transcription factors (TFs) and chromatin modifications such as enhancer-associated histone marks, which are typically localized at specific, narrow sites and have on the order of thousands of binding

sites, 20 million of reads may be adequate (4 million reads for worm and fly Tfs). Protein with more binding sites (e.g. RNA PolII) or broader factors, including most histone marks, will require more reads, up to 60 million for mammalian ChIP-seq. Importantly, to ensure sufficient coverage of a substantial portion of the genome and non-repetitive autosomal DNA regions, control samples have to be sequenced significantly deeper than ChIP samples especially in experiments analyzing TFs or involving diffused broad-domain chromatin data. To guarantee that the chosen sequencing depth is adequate, a saturation analysis is recommended, i.e., called peaks should be consistent when the two steps of read mapping and peak calling are performed on increasing numbers of reads chosen at random from the actual reads. A typical workflow for the computational analysis of ChIP-seq data consist in the following steps

i) read mapping and quality metrics;
ii) peak calling;
iii) iii) downstream analysis.

In the first step, before mapping the reads to the references genome, a quality cutoff has to be applied to filter out low quality sequences. Among the various methods, FastQC can be used for an overview of the data quality, while Phred quality scores can be used to describe the confidence of each base call in each sequence tag and to filter low quality reads. Sometimes, after this step is necessary trim the end of low-quality reads. Furthermore the library complexity can be measure using R custom script or the R *preseq* package. The low library complexity is linked to many factors such as antibody quality, over-cross-linking, amount of materials, sonication, or over-amplification by PCR. The latter can be corrected by systematic identification and removal of redundant reads, which is implemented in many peak callers as it many improve their specificity. The remaining reads should then be mapped using one of the available mappers as Bowtie, BWA (Li H. et al., 2010), SOAP (Li et al., 2008) or MAQ.

The second step, peak calling, is a pivotal analysis for ChIP-seq to predict the regions of the genome where the chipped protein is bound by finding region with significant numbers of mapped reads (peaks). Given a set of aligned sequences, peaks identified as genomic regions with enriched signals, i.e., where more sequences are aligned than would be expected by chance, thus indicating locations of binding sites or histone modifications. Several *peak-calling* algorithms that scan the genome to identify enriched regions are currently available. These algorithms take the advantage of the directionality of the reads. The fragments are sequenced at the 5'end and the locations of mapped reads should from two distributions, one on the positive strand and the other on the negative strand, with a consistent distance between the peaks of the distributions. In these methods, a smoothed profile of each strand is constructed and the combined profile is calculated either by shifting each distribution towards the center or by extending each mapped position into an appropriately oriented fragment and then adding the fragments together. The latter approach should results in a more accurate profile with respect to the width of the binding, but requires an estimate of the fragments size as well as the assumption that fragment size is uniform. Given a combined profile, peaks can be scored in several ways. A simple fold ratio of the signal for the ChIP sample relative to that of the control sample around the peak provide important information, but it is not adequate. A Poisson model for the tag distribution is an effective approach that accounts for the ratio as well as the absolute tag numbers, and that can also be modified to account for regional bias in tag density due to the chromatin structure, copy-number variation, or amplification biases. A binomial distribution or other models can also be used. A major difficulty in identifying enriched

regions is that peaks can present three forms, i.e. sharp, broad and mixed. In fact, the binding of transcriptional factors is governed by signal, in contrast the signals for histone modifications is usually diffuse and lack of well-defined peaks. Tools such MACS (Zhang et al., 2008) were developed to identify sharp peaks in contrast to SICER (Zhang et al., 2009) that identify broader regions. Instead, spp was developed to identify peaks with mixed profiles.
Finally, the downstream analysis includes differential binding analysis Deseq (Anders et al., 2010), edger (Robinson et al., 2010), peaks annotation, using GREAT (McLean et al., 2010), ChIPpeakAnno (Zhu et al., 2010), and motif analysis.

2.2.2 Public databases of ChIP-seq data

ChIP-seq data are mainly deposit in Gene Expression Omnibus (GEO), a public repository for a wide range of high-throughput experimental data. The organization of ChIP-seq data in the GEO database is similar to that of gene-expression microarray data. GEOP series (GSE) contain the lists of GEO samples (GSM) that together form a single experiment. The GSM file contains information that entirely describes an experiment. The main fields are the *title*, the *sample type*, which specifies the format file, uploaded on GEO (.sra, .bed, or .bam), the *source name* that describse the kind of tissue or cell lines used in the experiments (e.g. Human embryonic stem cells received from the James Thompson laboratory. Cells are from the H9 line.; renlab.H3K23me2.hESC.H9.02.01). The *description field* contains information about the design of the experiment (e.g. the kind of antibody used, the type of sequencing platform). Finally the *data processing field* contains descriptions about upstream analysis of raw sequences. The GSM page contains also the links to download the all data files. Usually the files of a ChIP-seq experiment are in the standard file format of next generation sequencing as .sra, .bam, or .bed. Often there is the necessity to download the .sra files and this step might be tedious especially when is required to recover multiple experiments. To circumvent this hurdle, we coded in R a script that automatically retrieves .sra files from a list of GSM. This code is based on two R packages, i.e., *GEOquery* and *SRAdb*. *GEOquery* is a bridge between GEO and the Bioconductor project and it was developed to download gene-expression microarray data in a R environment. *SRAdb* was developed to facilitate the access to the data on SRA database. The code works in three simply steps (Figure. 5):

1. download of meta-data associated with a GSM with *getGEO* functions;
2. recovery of the .srx from the meta-data;
3. creation of a directory on the local computer using as labels the GSM and others custom strings (e.g. version of genome, cell type);
4. creation of a list of .sra associated with a .srx using *listSRAfile*;
5. download of all .sra files with *getSRAfile* function using as parameters sraType=".sra".

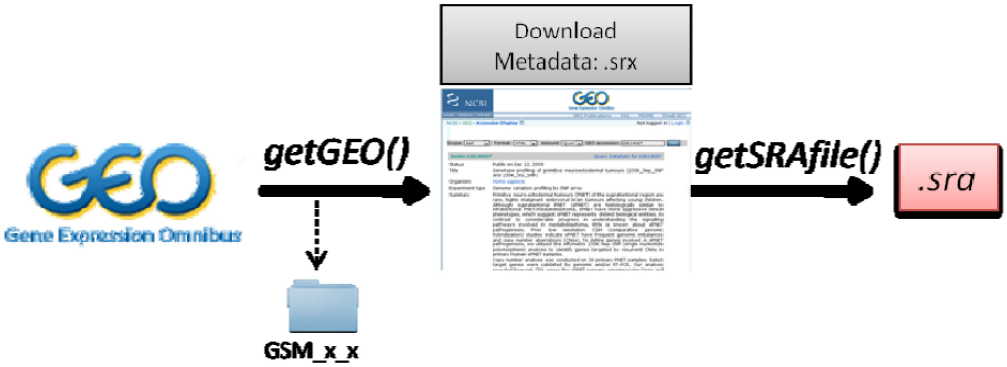

Figure 5. Workflow of the script to download ChIP-seq data

2.2.3 Reduction of data complexity using ChIP-seq data

Because ChIP-seq experiments produced million of short sequence reads, the *reduction of data complexity* is a mandatory step in the analysis of chromatin immuno-precipitation sequencing experiments. The simplest approach to achieve this objective is mapping the reads again a reference genome followed by peak calling. In this work, we designed a bash pipeline named *run-preprocessing.sh* that first aligns the reads on the reference genome and then performs peak-calling analysis using SICER, a peak-caller that was developed for histone-chip data. A similar version, using MACS to perform peak calling, was designed to analyze data for transcription factors. The pipeline coded in *bash* use others packages such as *sratoolkit2.1.7*, *bowtie0.12.7* (Langread et al., 2009), *samtools-0.1.18* (Li et al., 2010), and *bedtools-2.15.0* (Quinlan et al., 2010). In particular, *sratoolkit* extracts raw reads from .sra files in a more suitable file format for the subsequent analysis, i.e. the .fastq format containing all reads of the experiment. Then reads are aligned to a reference genome using *bowtie*. *Samtools* is used to generate high quality .bam files by removing the PCR clones with the modules *samtools rmdup*. Finally .bam files are converted into .bed files with *bedtools*. *Run-preproccesing.sh* was developed to use different input file as .sra, .fastq, .bam, while the output files of the analysis are always .bam and .bed. Finally, since *run-preprocessing.sh* can work using different input files, every step of processing is not mandatory, e.g. when the input is a .bam with already aligned reads the bowtie alignment step can be skipped. A simply version of *run-preproccesing.sh* code is reported in the following paragraphs along with an explanation of all parameters:

 # sratoolkit
sratoolkit.2.1.7-centos_linux64/fastq-dump .sra -O ./

 # alignment of reads
Bowtie -S -t -p 6 -m 1 hg19 file.fastq > ./file.sam 2> ./.bowtie.summary.txt
-S: print alignments in sam format
-p: number of processor
-m: mis matched
Hg19: reference sequence

 # conversion .sam to .bam
samtools view -S -h –b .sam -o bam
-S: the input is a sam file.
-h: include header in the output
-b: output in the bam format

 # sorting of bam files

samtools sort ./.bam ./file.bam.sort

#generate statistics of .bam sorted files
samtools flagstat ./file.bam.sort > ./file.bam.sort.stats

 # remove PCR duplicate
samtools rmdup -s ./file.bam.sort ./file.ready.bam 2> /dev/null
-s: remove duplicate for single end reads.

 # indexing bam files
samtools index ./file.ready.bam

 # statistics of bam without the duplicates reads
samtools flagstat ./file.ready.bam ./file.ready.bam.stats

 # create high quality .bam files
samtools view -b -h -F4 -q15 ./file.ready.bam > ./file.HQ.bam
-b: output in .bam format
-h: include the header in the output
-F4: skip alignments with bits present in INT [0]
-q: skip alignments with mapq smaller than INT [0]
 # indexing HQ bam file
samtools index $ffv_no_extension.HQ.bam

 # conversion of HQ bam file in bed file
BEDTools-Version-2.15.0/bin/bedtools bamtobed -i ./file.HQ.bam > ./file.HQ.bed
-i: input file

Peak-calling was performed using SICER. This method pools together all signals from the nucleosomes located together in the same of modification state. This feature improves the signal-to-noise ratio and is especially useful in difficult cases of diffuse enrichment covering extended genomic regions produced by histone modifications, for which enrichment at any short distance of one or several nucleosomes does not appear to be enough significant. The method involves scoring of each potential ChIP-enriched domain according to the collective profile of enrichment on the domain. SICER makes use of a mathematical model for the distribution of score in a genomic background model of random reads, and employs this theory to identify spatial clusters, large and small, unlikely to appear by chance. Using a control library, SICER identifies a set of candidate domains that exhibit clustering ChIP signal using the random background model, and compares the strength of the ChIP signal with that of the control signal at each candidate domain to determine the significance of enrichment (Equation 1). In particular, given a genome of length *L* partitioned in *w* non-overlapping windows, the score *s* for a window *w* with *l* reads is:

$$s(l) = -logP(l, \lambda) \quad (1)$$

where $P(l, \lambda)$ is a Poisson distribution parameterized by the average number of reads in a window $\lambda = wN/L$, with *N* being the total number of reads.
Each window is eligible if the read count in the window is equal to or above (below) a read-count threshold l_0. The l_0 is determined by a p-value requirement based on a Poisson distribution:

$$\sum_{l=l_0}^{\infty} P(l,\alpha) \leq p_0 \quad (2)$$

and l_0 depends on the ChIP-seq library size. The eligible windows are separated by gaps, which are regions of ineligible windows in between two eligible windows. The islands of clusters are identified as clusters of eligible windows separated by gaps (g) (Figure 6).

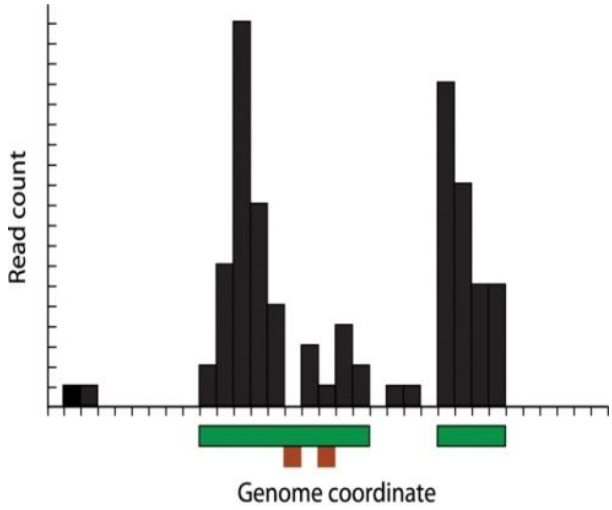

Figure 6. Schematic illustration of island definition. The x-axis reports the genome coordinates, where each interval represent a window. The y-axis denotes the read count. Regions underlined by the green horizontal bars are the two identified islands. The two underlined boxes are gaps in the first island

2.3 Cap Analysis of Gene Expression (CAGE)

Various technologies have been developed to deduce and quantify the transcriptome, including hybridization or sequence-based approaches. Hybridization-based approaches typically involve incubating fluorescently labeled cDNA with custom-made microarrays or commercial high-density oligo microarrays. Genomic tiling microarrays, that represent the genome at high density, have been constructed and allow the mapping of transcribed regions to a very high resolution, from several base pairs to ~100 bp. The hybridization-based methods have several limitations, which include reliance upon existing knowledge about genome sequence, high background levels owing to cross-hybridization, and a limited dynamic range of detection owing to both background and saturation of signals. Moreover, comparing expression levels across different experiments is often difficult and can require complicated normalization methods. In contrast to microarray methods, sequence-based approaches directly determine the cDNA sequence. Initially, Sanger sequencing of cDNA or EST libraries was used, but this approach is relatively low throughput, expensive and generally not quantitative. In recent years several technologies have become available to sequence the DNA at a very high throughput. Although these technologies have originally been used for genomic sequencing, more recently researches have turned to using deep sequencing or (ultra) high throughput technologies for a number of other applications. For example, several researches have used deep sequencing to map histone modifications genome-wide, or to map the locations at which transcription factors bind DNA. Another application that is rapidly gaining attention is the use of deep sequencing for transcriptome analysis through the mapping of RNA fragments. An alternative new high-throughput approach to gene expression analysis is *Cap Analysis of Gene Expression* (CAGE). CAGE is a

relatively new technology introduced in 2003 by Carninci (Carninci et al., 2003). Cap Analysis Gene Expression (CAGE) was introduced as method to determine transcription start sites on a genome-wide scale by isolating and sequencing short sequence tags originating from the 5' end of RNA transcripts. Mapping these tags back to the reference Genome identifies the transcription start sites from which the transcripts originated. CAGE relies on a cap-trapper system to capture full-length RNAs while avoiding rRNA and tRNA transcripts (Figure 7).

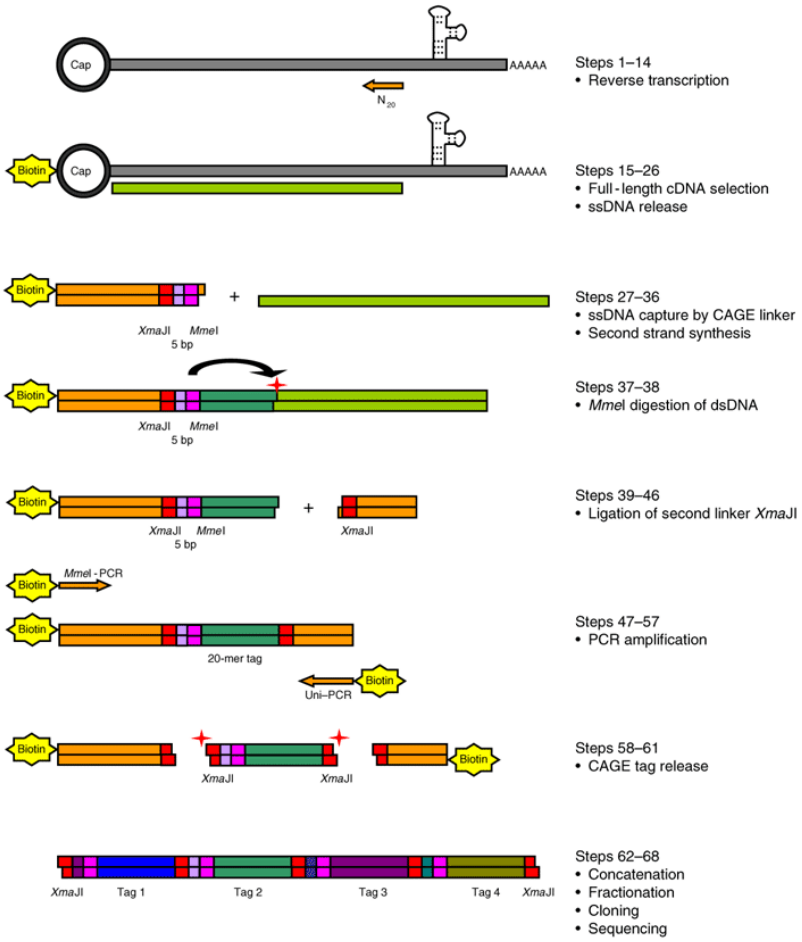

Figure 7. Cap Analysis of Gene Expression

First, an oligo-dT primer is used to reverse-transcribe poly-A terminated RNAs. Alternatively, a random primer can be used for RNAs without a poly-A tail, which may constitute almost half of the transcriptome. Subsequently, full-length cDNA are select by biotinylated cap-trapped their 5' cap structure, allowing capture by streptavidin-coated magnetic beads. Ligation of a specific linker sequence containing an *Mme1* recogninition site to the 5' end of the full-length cDNA creates a restriction site about 20 nucleotides downstream, producing a short CAGE tag starting at the 5' end of eukaryotic mRNAs. Then, 5' ends tags ('CAGE tags') are isolated and after the cleavage, sequences tags are purified and concatenated together with appropriate linkers for sequencing with the 454 Life Science sequencer. New sequencing platforms such as Illumina and SOLiD will use tags directly, without concatenation. At such sequencing depths significantly expressed TSSs are typically sequenced a large number of times. It thus becomes possible to not only map the locations of TSSs but also quantify the expression level of each individual TSS,

then CAGE it is a unique tool in the analysis of transcriptional regulatory networks. There are several advantages that deep-sequencing approaches to gene expression analysis offer over standard micro-array approaches. First, large-scale full-length cDNA sequencing efforts, have made it clear that most if not all genes are transcribed in different isoforms owing both to splice variation, alternative termination, and alternative TSSs. One of drawbacks of micro-array expression measurement has been that the expression measured by hybridization at individual probes is often a combination of expression of different transcript isoforms that may be associated with different promoters and may be regulated in different ways. In contrast, since deep sequencing allows measuring expression along the entire transcript, the expression of individual transcript isoforms can, in principle, be inferred. The first step in the analysis of deep-sequencing expression data is the mapping of the (short) reads to the genome from which they derive. CAGE tags are usually mapped to a reference genome using a novel alignment-algorithm called Kalign2 (Lassmann et al. 2012) that maps tags in multiple passes. Once the RNA sequence reads or CAGE tags have been mapped to the genome, is obtained a collection of positions for which at least one read/tag is present. In multiple samples could be obtain, for each position, a read-count or tag-count profile that counts the number of reads/tags from each sample, mapping to that position. These tag-count profiles quantify the expression. CAGE-tag based expression measurements directly link the expression to individual TSSs, thereby providing a much better guidance for analysis of the regulation of transcription initiation.

2.3.1 CAGE library preparation and sequencing

For the CAGE data analyzed in this work, RNA has been extracted using RNeasyPlus Mini kit as per the manufacturer instructions (QIAGEN). DeepCAGE library preparation was performed by DNAFORM Inc. at RIKEN Omics Science Center. Briefly, the cDNA synthesis was performed with 5 µg of total RNA, the random (N15) reverse transcription primers and PrimeScript Reverse Transcriptase (TAKARA). Capped RNA was biotinylated and treated with RnaseOne. Then, hybrid cDNA with biotinylated Capped RNA was selected with cap-trapper method. cDNA was released from streptavidin beads. A sample-specific linker, containing a recognition site for the sample-specific barcode sequence (3 bp) and the type III restriction-modification enzyme EcoP15I, was ligated to the single-strand cDNA. After ligation, the 2nd strand synthesis was performed and the resulting double-stranded cDNA was cleaved with EcoP15I. After, a second linker was ligated to the CAGE tag.The CAGE tags were separated from unmodified DNA with streptavidin beads. Then, the DNA fragments were PCR-amplified by using linker-specific primers. Samples were sequenced using the Illumina GAIIx sequencer. Each library was sequenced in one lane of a single strand 38 bp Illumina Run. DeepCAGE Data Analysis was performed by DNAFORM at RIKEN GeNAS. Briefly, tags were extracted and mapped to human genome version hg19 (NCBI build 37). Tags mapping the human ribosomal DNA sequence were eliminated. For CAGE tags mapping to multiple genome locations, a weighting strategy, based on the number of CAGE tags within a 200bp neighborhood around each candidate mapping location, was applied. Equal weights were used if no unique tags were found within the 200 bp region for all candidate mapping locations (Faulkner et al., 2008).

2.3.2 Normalization of CAGE data

The tag-count profiles quantify the expression of each position across samples and the

simplest assumption would be that the true expression in each sample is simply proportional to the corresponding tag-count. Recent papers, dealing with RNA-seq data simply count the number of read/tags per kilobase per million mapped reads/tags. Similarly, previous efforts in quantifying expression from CAGE data simply defined the *tags per million* (TPM) of a TSS as the number of CAGE tags multiplied by 1 million (Equation 3):

$$TPM = 100000x\left(\frac{absolute\ tag\ count}{total\ tag\ count}\right)$$
$$relative\ error = \frac{1}{\sqrt{absolute\ tag\ count}} \quad (3)$$

However, such simple approaches assume that there are no systematic variations between samples (which are not controlled by the experimenter) that may cause the absolute tag-counts to vary across experiments. However, in 2009 van Nimwegen and collaborators (2009) described a normalization method that takes in account the variations between samples. In particular, the noise in the expression measured across different CAGE replicated samples was modeled by a convolution of multiplicative noise and Poisson sampling to derive an analytical approximation to the resulting noise distribution. The noise model allows to rigorously assessing the statistical significance of measured expression differences across different samples. Systematic variations may result from the quality of the RNA, variation in library production, or even biases of the employed sequencing technology.

2.3.3 Clustering of CAGE tags and promotrome construction

Traditionally it has been assumed that each gene has a unique promoter and TSS, but large-scale sequence analysis, such that performed in the FANTOM3 project, made clear that most genes are transcribed in different isoforms that use different TSSs. Alternative TSSs not only involve initiation from different areas in the gene locus, but TSSs typically come in local clusters spanning regions ranging from a few to over 100 bps. These observations raise the question as to what an appropriate definition of basal promoter is. In the FANTOM3 researcher aimed at characterizing genomic regions that contain a significant amount of transcription initiation. To answers to this question, they clustered CAGE tags whose genomic mapping overlapped by at least 1bp. In another study, the identification of all regions in which the density of CAGE tags is higher than a given cut-off was performed using hierarchical clustering. Unfortunately, both methods cluster CAGE tags based only on the overall density of mapped tags along the genome ignoring the expression profile of the TSSs across different samples. However, it is clear that, to study transcriptional regulation, it is important to properly separate local clusters of TSSs that are co-regulated from those that show distinct expression profiles. As such, van Nimwegen et al. (2009) described a Bayesian clustering method that clusters nearby TSSs into *transcription start clusters* (TSCs) that are co-expressed in the sense that their expression profiles are statistically indistinguishable. For each significantly expressed TSC there will be a *proximal promoter region* that contains regulatory sites that control the rate of transcription initiation from the TSSs within the TSC. Since TSCs can occur close to each other on the genome, individual regulatory sites may sometimes be controlling multiple nearby TSCs. Therefore, in addition to clustering nearby TSSs that are co-expressed, they introduced a further

clustering layer, in which TSCs with overlapping proximal promoters are clustered into *transcription start regions* (TSRs). Then, whereas different TSSs may share regulatory sites, the regulatory sites around a TSR only control the TSSs within the TSR.

2.3.4 Annotation of CAGE peaks with known database

With the aim to annotate TSCs and TSRs, we created a custom R script that annotates promoters on the base of their vicinity to *RefSeq* gene annotated genes, *ENSEMBL ncRNA*, ncRNA included in publicly available data sets (Cabili et al., 2010; Khalil et al., 2009; Kretz et al., 2012), *Vertebrate Genome Annotation* (Vega), pseudogenes (Wilming et al., 2012), Yale Gerstein Group pseudo-genes (Zhang et al., 2006), and gencode (Harrow et al. 2006). This annotation contain 55889 genes, 20078 protein-coding genes, 12933, long non-coding RNA genes, 12933, small non-coding RNA genes, 9173 pseudo-genes, 190051 transcripts, 80413 protein-coding transcripts, 12421 nonsense mediated decay transcripts, and 21271 long non-coding RNA loci transcripts. The Cabili dataset contains a catalog of 8195 human lincRNAs found integrating RNA-seq data from 24 tissues and cell types with publicly available transcript annotations. The Khalil dataset is a collection of 1703 large non-coding RNAs obtained using distinctive chromatin signature that mark actively transcribed genes. The signature consists of a short region with histone H3 lysine 4 trimethylation (H3K4me3) (corresponding to the promoter) and a longer region with histone H3 lysine 36 trimethylation (H3K36me3, corresponding the transcribed region). Finally, the Kretz dataset contains 836 long noncoding RNAs (lncRNAs) detected with RNA-seq experiments, chromatin signature of H3K4me3 and H3K36me3 and tiling array. For each database, the annotation with the smallest distance to the CAGE-defined promoter on the same chromosome strand was found. To define the distance to the CAGE-defined promoter we reasoned as follows:

1. if the 3' end of the promoter is upstream of the 5' end of the annotation, then the distance corresponds to that between the 3' end of the promoter and the 5' end of the annotation;
2. if the 5' end of the promoter is downstream of the 5' end of the annotation, then the distance corresponds to that between the 5' of the annotation and the 5' end of the promoter;
3. otherwise, if the promoter overlaps the 5' end of the annotation, then the distance is zero.

If this distance is less than 400 bp, then the promoter is associated to the gene or transcript.

2.4 Prediction of cis-regulatory elements using the Intersect Reads Count Method (IRCM)

To identify cis-regulatory elements from ChIP-seq data downloaded from public repositories and in house experiments, we developed a pipeline called intersect-read-counts-method (IRCM). This procedure has been designed to handle datasets of ChIP-seq with only two histone modifications, i.e. H3K4me1 and H3K4me3, as H3K4me1 and H3K4me3 compose the simplest combination of histone marks characterizing promoters and enhancers. In the genome, the presence of broader regions of H3K4me3 with low levels of H3K4me1 characterizes active promoters, in contrast enhancers have an opposite pattern.

The pipeline comprises three steps and uses as input the files generated from the reduction of complexity analysis performed with SICER. These files (.bed) contain the coordinates on the genome of the significant regions of H3K4me1 and H3K4me3 and the quantification of signals in correspondence of them. The quantification of signal in a particular genomic region is quantified by SICER using the criteria described in the paragraph 2.2.3. Briefly, a estimate of the number of reads in a region is first computed and, then, this signal is fitted with a Poisson distribution. Specifically,

1) in the first step, the script invokes BEDtools to identify regions where H3K4me1 peaks overlap (do not overlap) with H3K4me3;
2) the second step considers only the overlapping regions, for which log-ratios between of H3K4me1 and H3K4me3 tag counts are quantified. The tag counts of H3K4me3 and H3K4me1 are first normalized using the sequencing depths of both libraries. Then, if the ratio is greater than 0, the region is label as *promoter*, otherwise as *enhancer*;
3) in the final step, the output of CAGE and ChIP-seq peaks are merged. The R script classifies CAGE peaks as mono-directionally transcribed or bi-directionally transcribed. Given two peaks, the script calculates the distance between them and label as bi-directionally transcribed those regions where the peaks lay at a distance of 1 kb, and mono-directionally transcribed otherwise. Finally, CAGE regions are merged with ChIP-seq peaks using ChIPpeakAnno, considering a distance of 2000 kb between a ChIP-seq and CAGE peaks.

2.5 Shannon entropy

To quantitatively measure the relative occupancy of each cis-regulatory element, we adopted a strategy based on Shannon entropy to assign a tissue-specificity index (Barrera et al., 2008) Specifically, for cis-regulatory elements, its relative occupancy in a tissue *t* is defined as:

$$\frac{B_{t,s}}{\sum_{t=1}^{N} B_{t,s}} \tag{4}$$

where $B_{t,s}$ is calculated as the normalized binding intensity relative to input and library size, N is the number of tissues, and the entropy score is defined as:

$$H_s = -1 \times \sum_{t=1}^{N} p_{t,s} \times log_2(p_{t,s}) \tag{5}$$

where the value of H_s ranges between 0 and $\log_2(N)$. An entropy score close to zero indicates the occupancy at this site is highly tissue-specific, while an entropy score close to log2(N) means the site is bound uniformly. This procedure was coded in R using the *entropy* package (Hausser et al., 2009).

2.6 Support vector machine

Support vector machines (SVMs) are a set of supervised learning algorithms invented by Vladimir Vapnik in 1995, at Bell AT&T laboratories. Support vector machines have a broadly applicability to many types of patterns recognition problems. SVMs rely on pre-

processing the data to represent patterns in a high dimension typically much higher than the original feature space. With appropriate non-linear and linear mapping to a sufficiently high dimension, data from two categories can be separated by a hyper plane. This choice is often be informed by the designer's knowledge of the problem domain. In absence of such information, one might choose to use polynomials, Gaussians or other basic functions. The dimensionality of the mapped space can be arbitrarily high. Defining the margin as any positive distance from decision hyper plane, the goal in training support vector machines is to find the separating hyper plane with the largest margin. The support vectors are the (transformed) training patterns that are (equally) close to the hyper plane (Figure 8). The support vectors are the training samples that define the optimal separating hyper plane and are the most difficult patterns to classify. Informally speaking, they are the most informative patterns for the classification task. The problem of minimizing the magnitude of the weight vector constrained by the separation can be reformulated into an unconstrained problem by the methods of Lagrange undetermined multipliers. Using the so-called Kuhn-Tucker construction, this optimization can be rewritten as a maximizing problem that can be solved using a quadratic programming.

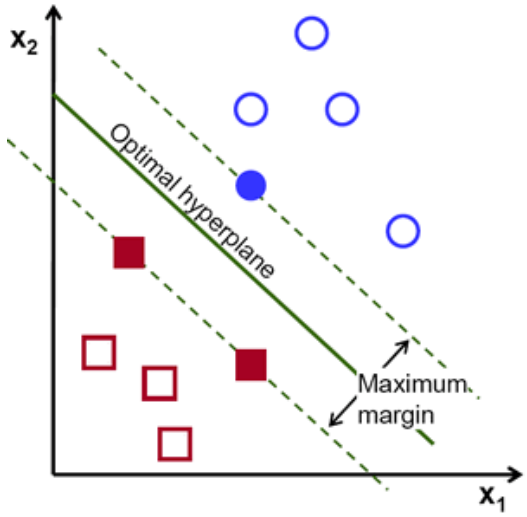

Figure 8. Three support vectors of a support vector machine classifier

A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one target value (i.e. the class labels) and several attributes (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data), which predicts the target values of the test data given only the test data, attributes. Given a training set of instance-label pairs (x_i,x_j), i=1,…,l where $x_i \in R^n$ and $y_i \in \{1,-1\}^l$, the support vector machines (SVM) (Boser et al., 1992; Cortes and Vapnik, 1995) require the solution of the following optimization problem:

$$\underset{w,b,\varepsilon}{min}\frac{1}{2}w^T w + C\sum_{i=1}^{l}\varepsilon_i \qquad (6)$$

subjected to $y_i(w^T\emptyset(x_i) + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0$

Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function $\emptyset$. SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. C > 0 is the penalty parameters of the error term. Furthermore,

$K(x_i, x_j) \equiv \emptyset(x_i)^T \emptyset(x_j)$ is called the *kernel function*. The basic kernel functions are:

- linear: $K(x_i x_j) = (= x_i^T x_j)$
- polynomial: $K(x_i x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$
- radial basis function (RBF): $K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2, \gamma\right) > 0$
- sigmoid: $K(x_i, x_j) = \tanh(\gamma\, x_i^T x_j + r)$.

2.6.1 Implementation of SVM2CRM

SVM2CRM is an R package for the genome wide-identification of cis-*regulatory elements* from chromatin immuno-precipitation data ChIP-seq. The program implement *LiblineaR* (Fan et al., 2008) a library that allows performing linear classification on large dataset with millions of instances and features. *LiblineaR* support L2-regularized classifiers (such a L2-loss linear SVM, L1-loss linear SVM, and logistic regression), as well as L1-regularized classifiers (such as L2-loss linear SVM and logistic regression). SVM2CRM has been developed taking account the modularity of the framework analysis. Therefore, it can be easily extended if custom analytical pipelines are required for more complex or specialized purposes, or to integrate *SVM2CRM* analysis into other computational pipelines. Moreover, since some steps are time consuming due to the repeated cycles of signal processing, some functions has been implemented to distribute the computations on several processors. The parallelization is based on *multicore* (Simon et al., 2009), a package that allows parallelizing R code using multiple cores and CPUs.

2.6.7 Framework of SVM2CRM analysis

The core of SVM2CRM procedure consists of four several steps:
1. data pre-processing;
2. tuning of SVM parameters;
3. prediction;
4. result storage (Figure 9).

During the pre-processing step, SVM2CRM exploits different functions to accommodate several input objects in relation to the kind of analysis to perform. The analysis can start using two input file, i.e., a reference with the positions of enhancers and promoters (e.g. .bed file) and the signals of histone modifications. Moreover, for more advance analysis the user can create two R objects using kmeans algorithm that contain the signals of histone modification associated with promoters and enhancers and clustered to reduce the complexity of signals. During the pre-preprocessing step one of the most important function is *cisREfind.* This function uses a sliding windows approach with user-defined parameters (size of windows, step) and processes the signals in three different ways: median, maximum, and shape function. Furthermore cisREfind can estimate, using the Pearson correlation function (*cor*), the correlation between a signal inside a window on the genome and reference signals of histone marks associated with cis-regulatory elements (for e.g. the output of kmeans). This could be useful when the user wants to perform profile matching.

In the second step, the use of tuningParametersComb in combination with performanceSVM allows to compute the performance of predictions considering the

different weights of two classes (e.g. enhancer, not enhancers), the kernels, the cost functions and others parameters of SVM. performanceSVM compute different parameters such as sensitivity, specificity, fscore, and false positive rate. The finally step returns the predictions with the best training model and creates a .bed output file with a list of predicted enhancers.

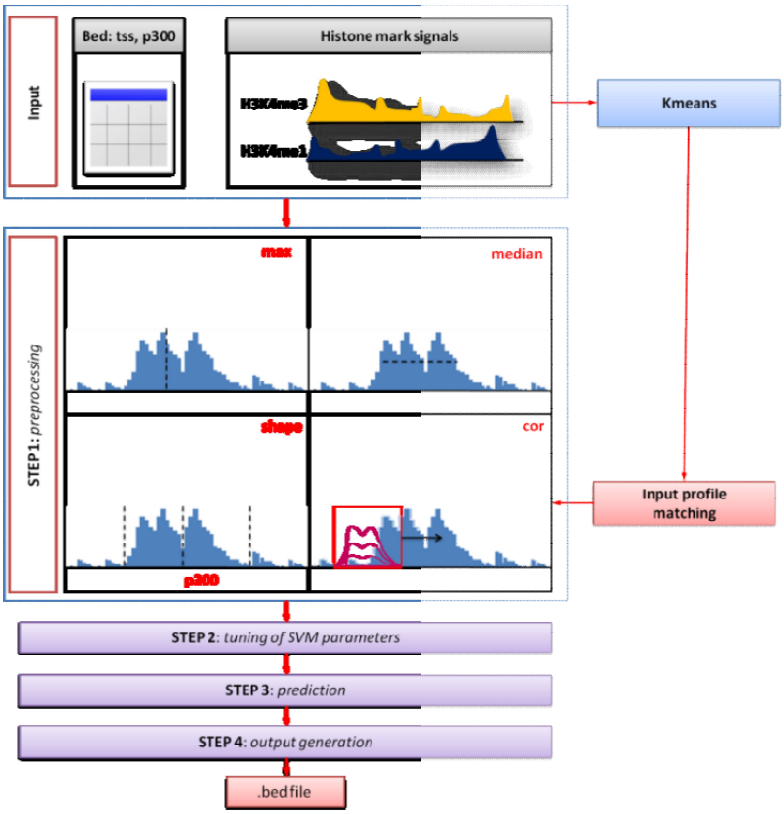

Figure 9. The SVM2CRM workflow

2.6.8 Identification of cis-regulatory elements using SVM2CRM

Before the analysis with SVM2CRM, ChIP-seq data have been pre-processed using *spp* R package (Kharchenko et al., 2008). ChIP-seq signals have been smoothed using a bandwidth of 200 bp and a step of 100. Epigenetics profiles associated with cis-regulatory elements and not-cis regulatory elements (derived from randomly selected regions) in chromosomes 1 and 2 have been used as positive and negative examples, respectively, to train different SVM models using a number of histone marks ranging from 2 to 8. The computational performances have been optimized using *liblinear*, a library of kernel independent, linear classifiers specifically indicated for the analysis of large datasets. As previously described, the SVM2CRM package exploits functions from the *multicore* package to perform parallel computation and reduce the computational costs. In the pre-processing step, the package uses two inputs, i.e., a reference file, with the positions of enhancers and promoters (e.g. bed file), and the histone modification signals. To reduce the complexity of the input data, histone signals can be first clustered using k-means clustering. During the pre-preprocessing step, histone mark signals (or clusters) are investigated, using a sliding window approach with user defined parameters (size of windows, step), to quantitatively characterize the various peaks. Specifically, any peak can be described using its median or maximum value or by its shape (*cisREfind* function). In addition, if some cis-regulatory elements are known, then the *cisREfind* function can implement the input profile matching

(Heintzman et al., 2007), i.e., can calculate the Pearson correlation between any signal inside a window and the reference signals of histone marks associated with the set of known cis-regulatory elements. In the tuning of SVM parameters, two functions (*tuningParametersComb* and *performanceSVM*) quantify the prediction performance in terms of sensitivity, specificity, fscore (true over false positives), and false positive rate of models derived from all combinations of histone marks, distributions of examples (enhancers and random regions), kernel types, cost functions, and other parameters. The metrics to estimate the performance of predictions with *performanceSVM* are compute as follow:

$$
\begin{aligned}
Sensitivity &= \frac{TP}{TP + FN} \\
Specificity &= \frac{TN}{TN + FP} \\
Precision &= \frac{TP}{TP + FP} \\
F - score &= \frac{2 \times Specificity \times Sensitivity}{Specificity + Sensitivity}
\end{aligned}
\tag{7}
$$

Results from tuning step allows selecting the best performing model that can, in turn, be used for predicting cis-regulatory elements (prediction of regulatory regions). Finally, SVM2CRM outputs the prediction in terms of a .bed file (generation of output step).

2.7 Experimental models

In this work we analyzed several different experimental models and datasets. Specifically, we considered experiments of ChIP-seq and Cap Analysis of Gene Expression (CAGE) produced at the Centre for Regenerative Medicine "Stefano Ferrari" of the University of Modena and Reggio Emilia Emilia for different cellular models of differentiation, i.e., hematopoiesis, human neural commitment, and keratinocyte commitment. In particular, ChIP-seq experiments used antibodies for H3K4me1 and H3K4me3 in cord blood derived hematopoietic stem cells $CD34^+/CD133^+$ cells, erythroid progenitors $CD36^+$, and myeloid progenitors $CD13^+$cells. The same antibodies were used on embryonic stem cells and neural stem cells progenitors, keratinocyte stem cells, and differentiated keratinocytes. Moreover, to identify enhancers using a optimal combination of histone marks, we analyzed a public-dataset of 8 distinct histone modifications: H3K4me1, H3K4me2, H3K4me3, H3K27me3, H4K20me1, H3K36me3, H3K27ac, H3K9ac in in embryonic stem cells (H1 ES), erythrocytic leukemia cells (K562), B- lymbhoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblast (NHLF), normal epidermal keratinocytes (NHEK), and mammary epithelial cells (HMEC). Data for chromatin states have been downloaded from ENCODE and Gene Expression Omnibus (GSE26386) databases. Data of p300 and DNAseI have been downloaded from ENCODE (Table 2).

Table 2. ChIP-seq data and data sources

Sample name	Data source	H3K4me1	H3k4me2	H3K4me3	H3k9me1	H3k9me3	H3k27me1	H3k27me3	H3k36me3	H4k20me1	H3k27ac	H3k9ac	H2AZ	PolII	Control
H1	*GEO*	*GSM466739*		*GSM469971*											*GSM605333*
H1	*GEO*	*GSM605312*		*GSM605315*											*GSM605335*
H1_der	*GEO*	*GSM818039*		*GSM767350*											*GSM767355*
H1_der	*GEO*	*GSM818040*		*GSM767351*											*GSM767356*
H9	*GEO*	*GSM667626*		*GSM605316*											*GSM667643*
H9	*GEO*	*GSM706071*		*GSM616128*											*GSM706081*
NES-cells	*UniMORE*														
KSC-cells	*UniMORE*														
NHEK_diff	*GEO*	*GSM733698*		*GSM733720*											*GSM733740*
CD133-cells	*UniMORE*														
CD36-cells	*UniMORE*														
CD34-cells	*GEO*														
MYE-cells	*UniMORE*														
ERY-cells	*UniMORE*														
Eryhtroblast_A	*Orkin et al.*	*GSM908035*		*GSM908039*											*GSM908073*
Erythroblast_F	*Orkin et al.*	*GSM908034*		*GSM908038*											*GSM908072*
H1	*ENCODE*	*GSM646341*	*GSM646343*	*GSM646345*				*GSM646337*	*GSM646339*	*GSM646349*	*GSM646336*	*GSM646347*			*GSM646351*
K562	*ENCODE*	*GSM646440*	*GSM646442*	*GSM646444*				*GSM646436*	*GSM646438*	*GSM646450*	*GSM646434*	*GSM646446*			*GSM646332*
GM12878	*ENCODE*	*GSM646322*	*GSM646324*	*GSM646326*				*GSM646316*	*GSM646320*	*GSM646330*	*GSM646316*	*GSM646328*			*GSM646332*
HepG2	*ENCODE*	*GSM646361*	*GSM646362*	*GSM646364*				*GSM646357*	*GSM646359*	*GSM646368*	*GSM646355*	*GSM646366*			*GSM646370*
CD133	*Cui et al.*	*GSM317586*		*GSM317587*	*GSM317588*	*GSM317589*	*GSM317583*	*GSM317584*	*GSM317585*	*GSM317590*			*GSM317582*	*GSM317591*	*no-control*
CD36-cells	*GEO- Cui et al.*	*GSM317596*		*GSM317597*	*GSM317598*	*GSM317599*	*GSM317593*	*GSM317594*	*GSM317595*	*GSM317600*			*GSM317592*	*GSM317601*	*no-control*

2.7.1 The hematopoiesis model

Hematopoiesis is a term used to describe the process of blood cell formation during both the embryonic and adult stages of an organism. Hematopoiesis is consider as the process of development, self-renewal and differentiation of hematopoietic stem cells (HSCs), a type of adult stem cell that is the source of all blood cell lineages. In fetal and adult mammals, HSCs predominantly reside in fetal liver and bone marrow, respectively. However, HSCs do not originate in fetal liver or bone marrow, but rather migrate from other tissues to these sites during embryonic development. Hematopoiesis is usually depicted in a hierarchical fashion, with HSCs giving rise first to progenitors and then to precursors with varying commitments to multiple or single pathways (Figure 10). Although this representation oversimplifies hematopoietic development, it provides a useful framework in which to consider the properties of the intermediate cell populations. In this hierarchical fashion hematopoietic cells can be broadly classified as transiting through three compartments: stem cell compartment, committed progenitor cells, and precursors of mature functional-end cells. The cells in each succeeding compartment are the progeny of, and more numerous than, the cells of the preceding compartment. The stem cell compartment is composed of very rare cells endowed with the ability to both self-renew and provide multi-lineage hematopoiesis for the life of the organism. Lineage markers are absent from these cells, and they normally are found in a quiescent state or are turning over very slowly. Stem cells are also equipped with a regimen of critical transcription factors that are important in the execution of their fundamental cellular functions of cell renewal and multi-lineage differentiation. The progenitor cell compartment contains cells that are found at a higher frequency than the stem cell pool and, like the stem cell, are not morphologically distinguishable. Their existence is revealed by their ability to give rise to differentiated progeny *in vitro* in well-defined functional assays. The progenitor cell compartment is derived from stem cells through a process of commitment to different lineage pathways. Transition of stem cells to cells of the committed compartment is achieved not by acquisition of new characteristics or new proteins but by enhancement of certain molecular pathways, already primed in these cells, and abrogation of others. As progenitor cells differentiate, they acquire more distinctive features characteristic of each lineage and move away from shared primitive progenitor characteristics: they show the enhancement of lineage-specific features, with a diminished or absence of expression of multi-lineage properties. In this manner, precursors for the various lineages arise. This precursor cell compartment is defined by morphological criteria and contains cells at different maturation stages; cell-surface and other markers can further distinguish these precursors. Precursor cells for each lineage follow a unique maturation sequence. These cells may be capable of undergoing a few rounds of cell division, or they may be end-stage, non-mitotic cells with a finite life span. The morphological characteristics of these cells reflect the accumulation of lineage-specific proteins, and organelles and the decline of nuclear activity, which gives them a unique appearance. A cellular roadmap that specifies lineage relationships between stem, progenitor, precursor and mature cells is indispensable for a comprehensive view of the transcriptional and epigenetic mechanisms that control normal development. In this regard the first comprehensive classical model of hematopoiesis was formulated. The first key postulate of this model is that loss of self-renewal capacity during differentiation precedes lineage commitment. Another crucial postulate of the classical model is that the earliest commitment decision segregates lymphoid and myeloid lineages. Lastly, the classical model predicts that lineage decisions occur as stepwise bifurcations. The immediate

progeny of HSCs are multipotent progenitor cells that retain full lineage potential yet have a limited capacity of self-renewal. Multipotent progenitors (MPPs) in turn give rise to oligopotent progenitors, which possess more restricted developmental potential: the common myeloid progenitors (CMPs) and the common lymphoid progenitors (CLPs), and each of these undergo further commitment steps (Figure. 10). CMPs give rise to the granulocyte/macrophage progenitors (GMPs), which become committed to the granulocyte-monocyte fate, and the megakaryocyte/erythroid progenitors (MEPs), which only produce erythroid and megakaryocyte (E-MK) cells. On the lymphoid side, CLPs give rise to B cell precursors and the earliest thymic progenitors (ETPs) committed to the T and NK lineages. The classical model is a simple yet powerful template for understanding blood development and interpreting the function of molecular regulators.

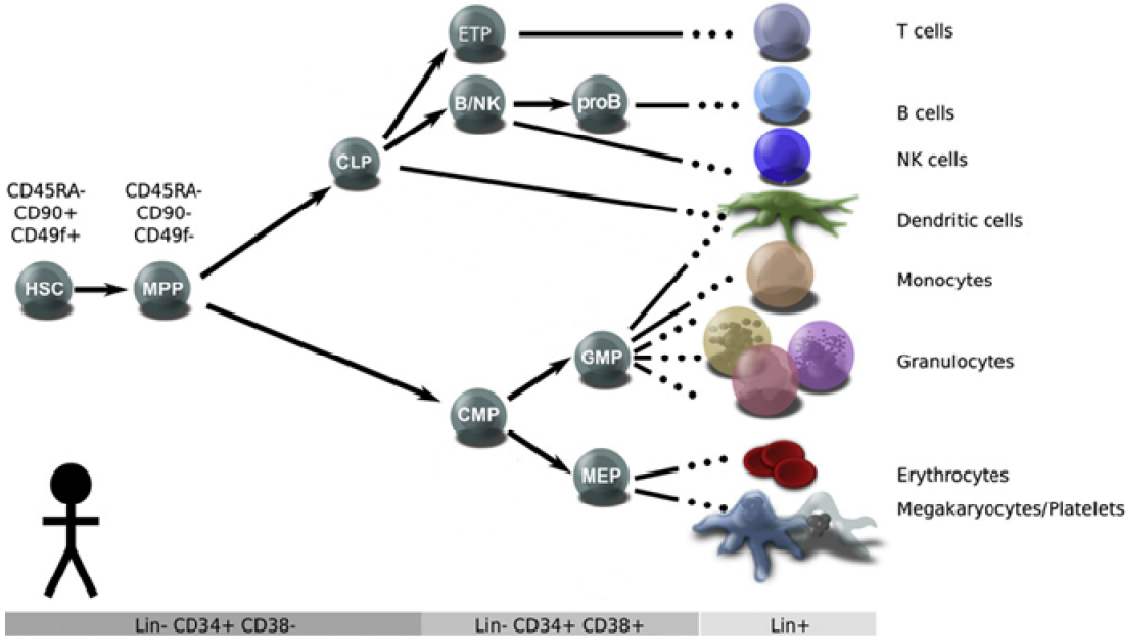

Figure 10. Hematophoiesis

2.7.2 Hematopoietic stem cells

Hematopoietic stem cell (HSC) is operationally defined as a cell that upon transplantation to an appropriately conditioned recipient reconstitutes the entire hematopoietic system (Orkin et al., 2000). HSCs are rare cells in the bone marrow and have two main properties. First, they can generate more HSCs, a process of self-renewal. Second, they have the potential to differentiate into various progenitor cells that eventually commit to further maturation along specific pathways. The end result of these events is the continuous production of sufficient, but not excessive, numbers of cells of all lineages (red and white blood cells, megakaryocytes and lymphoid cells).

2.7.3 Phenotypic characterization of hematopoietic stem cells

A major obstacle to studying HSC biology is that the cells are extremely rare. Only 1 in 10^6 cells in human bone marrow is a transplantable HSC (Wang et al., 1997), requiring purification from the bulk of differentiated cells. Purification of human HSCs requires simultaneous detection of several independent cell surface markers, so human HSCs and progenitors isolation has been characterized by enriching a rare cell population with a

combination of monoclonal antibodies. Over the past decade, various cell surface and metabolic markers have been identified and used to isolate human HSCs and progenitors. Among them, the CD34 antigen was the first marker found to enrich human HSCs and progenitors and it became the most critical cell surface marker to define the phenotype of human HSCs. Moreover of the most widely used associated antigen is CD38, as marker of more differentiated progenitor that negatively enrich for HSCs (Bathia et al., 2007). Cells that do not express, or express low levels of CD38 in association with high levels of CD34 ($CD34^{+}CD38^{-/low}$), constitute a widely accepted phenotype for identification of primitive HSCs. Finally another important marker of HSCs is CD133 (also known as PROM1). In conclusion, it is clear that the HSC has a wide range of phenotypes, however, the $CD34^{+}$ $CD90^{+}$ $CD38^{-}$ Lin^{-} designation has emerged over the past decade as a picture that identifies human HSCs in fetal liver and bone marrow, umbilical cord blood, adult bone marrow and mobilized peripheral blood (Doulatov et al., 2007).

2.7.4 Multipotent hematopoietic progenitors

$CD34^{+}$ cells are a heterogeneous population covering not only stem cells but also earlier multipotent progenitors and later lineage-restricted progenitors. In human hematopoiesis, populations enriched for HSC activity have been identified, but multipotent progenitor activity has not been uniquely isolated. There is reason to believe that cells defined as human HSCs by the cell surface markers previously described are likely to include one or more multipotent progenitor populations.

2.7.5 The erythropoiesis model

Erythropoiesis is a multi-step process involving commitment, proliferation, and differentiation of hematopoietic progenitors to mature, terminally differentiated erythrocytes (Figure 11). Acquisition of the red blood cell phenotype involves coordinate expression of regulatory and structural proteins acting in concert to direct development of progenitor cells into mature erythrocytes. The earliest erythroid progenitor, the BFU-E (burst forming unit-erythroid) expresses the cell surface antigen, CD34, as do all other early hematopoietic progenitors allowing for its isolation using anti-CD34 antibodies. The stage after the BFU-E, right before hemoglobin production begins is the CFU-E (colony forming unit-erythroid). BFU-Es and CFU-Es both express the CD36 antigen. CFU-Es express also high level of CD71 antigen (the transferrin receptor), which then decreases slightly during further maturation.

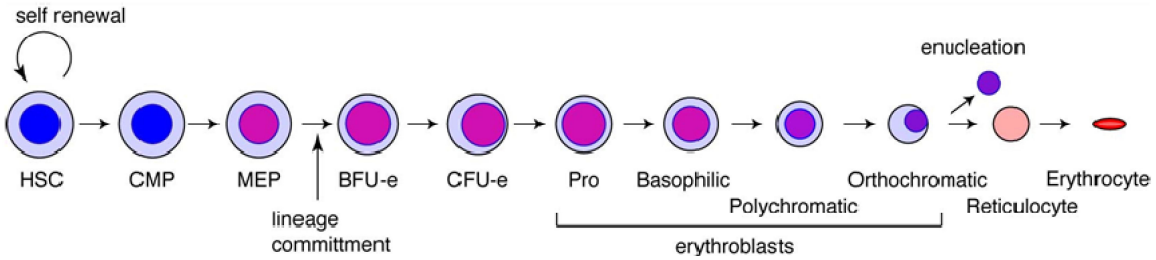

Figure 11. Commitment and differentiation of hematopoietic progenitors to mature, terminally differentiated erythrocytes. Abbreviations: HSC, hematopoietic stem cell; CMP, common myeloid progenitor; MEP, megakaryocytes/erythrocyte progenitor; BFU-E, burst forming unit-erythroid; CFU-E, colony forming unit-eryhtroid

2.7.6 The myelopoiesis model

The development of myeloid cells, consisting of granulocytes (neutrophils, eosinophils, basophils) and cells of monocyte/macrophage lineage, from HSCs involves a step-wise progression of progenitor cells in the bone marrow (Figure 12). Successive commitment steps include common myeloid progenitors (CMPs) and then granulocyte-macrophage precursors (GMPs), which can form CFU-GM (colony forming unit-granulocyte, monocyte) *in vitro*. GMPs are $CD34^+$ cells that are committed to differentiate into myeloid cells and can be commonly recognized by the surface expression of CD13 and CD33, occurring very early during myeloid cell maturation. In the next step, CFU-GMs may give rise to CFU-G (colony forming unit-granulocyte) and CFU-M (colony forming unit-monocyte). CFU-Gs give rise to myeloblasts, which are precursors that finally differentiate in neutrophils, eosinophils and basophils; while CFU-Ms give rise to monoblasts, which finally differentiate in monocytes.

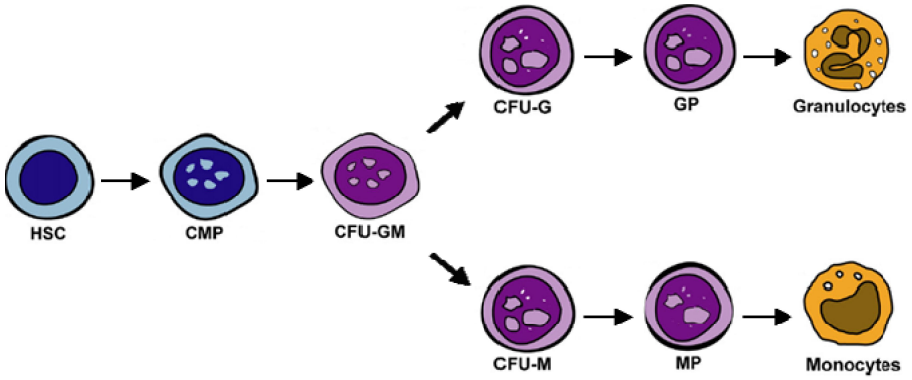

Figure 12. Commitment and differentiation in myelopoiesis. Abbreviations: HSC, hematopoietic stem cell; CMP, common myeloid progenitor; MEP, megakaryocytes/erythrocyte progenitor; BFU-E, burst forming unit-erythroid; CFU-E, colony forming unit-eryhtroid.

2.7.8 Chromatin signatures in human hematopoietic cells

Although in recent years different pathways and genes involved in hematopoiesis have been identified it remains unclear how the balance between self-renewal and differentiation is controlled and how a decision for differentiation is specified at molecular levels. It is clear that transcription programs, which include both activation of genes involved in the target lineage and repression of genes involved in non-target lineages, play essential roles during this process of fate determination. These specific transcription programs are controlled by a close coordination between transcription factors and chromatin states, both of which are regulated by extracellular signals. Appropriate chromatin modifications, including histone modifications, can help to maintain a relatively stable expression pattern of either activation or repression in stem cells or terminally differentiated cells. Bivalent domains (associated with both H3K4me3 and H3K27me3 modifications) have been found associated with many critical regions involved in pluripotency and differentiation of ESCs. These bivalent domains have been proposed to play critical roles and can be resolved to monovalent modification in ESCs differentiation. Recently, Cui et al. have examined the genome-wide distribution of chromatin modifications in human hematopoietic stem/progenitor cells (HSPCs) and in differentiated erythrocyte precursor cells. They showed that differentiation of $CD133^+$ cells into $CD36^+$ cells is accompanied by dramatic changes of histone modifications at critical genetic regions. Their data indicate that numerous genes involved

in the development and differentiation of multiple lineages are associated with both the repressive mark H3K27me3 and the active mark H3K4me3 in HSPCs. The 19% of the bivalent genes lost H3K27me3 and became activated after differentiation into $CD36^+$ cells. Interestingly, a large fraction of these genes were associated with H3K4me1 and H3K9me1 modifications in the HSPCs stage and also bound with RNA Pol II, suggesting that they were poised for activation. In contrast, the association of these mono methylations with the bivalent promoters, which lost H3K4me3 after differentiation (53%), was decreased dramatically in the HSPCs stage; in addition, almost no Pol II was detected in these promoters. These results are consistent with a model in which the fate of bivalent genes during differentiation is controlled by epigenetic modifications including various mono methylations that occur in the HSPCs stage. The mono methylations are correlated with the basal level expression of a subset of differentiation genes in HSPCs and therefore may be involved in the maintenance of their activation potential required for differentiation. Moreover, these studies showed that the differentiation into $CD36^+$ cells is accompanied by a genome-wide silencing of the genes involved in the development and differentiation of other lineages and that the silent states are locked off by a switch of epigenetic modifications.

2.7.9 Human neural commitment

Embryonic stem cells (ESCs) are a promising source of neural derivatives that can be used in stem cell-based therapies for treating genetic and acquired neurological disorders. The derivation of specific neuronal or glial cell types from ESCs invariably includes the production of neural stem cells, possibly multipotent, phenotypically stable and easy growing in culture. Several protocols used to neutralize ESC attempt to *in vitro* recapitulate the intermediate stages of the transition from pluripotent to neural stem/precursor cells. Recent data have described how specific growth factors and signaling molecules drive ESC neural induction and neural identity acquisition, a process marked by the down-regulation of pluripotency markers such as OCT4 and NANOG, while neuroectoderm markers, including NES and SOX1, are up-regulated. However, a deep knowledge of the molecular and regulatory circuitries driving human cell-fate restriction and in particular neural commitment is still lacking, and the identification of the genomic regions involved in this process still represents a great challenge. Two main approaches have been developed to neutralize ESCs, the first mimicking the environment that produces neuroectoderm in the embryo, by providing appropriate cell-cell interactions and signals through embryoid body formation, the second depriving the ESCs of both physiological cell interactions and molecular signals by low density culture in serum-free condition, evoking a default mechanism for ESCs neural induction and differentiation. Some protocols combine aspects of both approaches, promoting cell interactions to facilitate formation of all three primary germ layers followed by neural lineage-specific selection under defined conditions. Independently from the method, neural induction and differentiation involves multiple stages that roughly mimic the neural specification during embryogenesis: primitive ectoderm, definitive ectoderm, neuroectoderm. The neuroectoderm-like cells are comparable to the primitive NSCs (Tropepe et al., 2001), expressing SOX1 and NES, and showing dependence to Leukemia Inhibitory Factor (LIF), a cytokine supporting the self-renewal of cultured embryonic stem cells (Patterson and Nawa, 1993). These cells maturation into NSCs is associated to a morphological change, as cells elongate into a radial glia-like phenotype, forming rosettes in monolayer. These neural rosettes express high levels of NES and SOX1, are LIF independent and responsive to FGF2 or EGF. These

precursors are much more committed and less multipotent than earlier SOX1+-neuroepithelial stem cells, showing a strong resistance to induced regionalization (Elkabetz et al., 2008). Conversely, self-renewing neuroepithelial stem/progenitor cells (NESCs) show features of multipotent neural precursors appearing at early stages of neural tube formation, having intermediate characteristics between neural rosettes and later radial glial progenitors. Indeed, NESCs retain a stable SOX2+/SOX1+ phenotype in long-term culture, and show a huge differentiation potential towards neuronal and glial fates. Very few is known about the changes in the regulatory layout, based on differential usage of regulatory elements, occurring at ESCs progressive loss of pluripotency in favor of neural-fate acquisition, make evident by the different features of the diverse neural stem/progenitor cells. Epigenetic changes occurring at putative enhancer regions during human ESCs induction to neuroectodermal spheres (NECs), non-homogeneous aggregates of neural stem/precursor and variably differentiated neural cells, were recently described (Rada-Iglesias et al., 2011). Enhancers were defined as genomic regions bearing several histone modifications (that are, H3K4me1, H3K4me3, H3K27me3 and H3K27ac) to distinguish distal elements from proximal promoters, and binding of chromatin regulators (that is, p300, BRG1). H3K4 methylation is known to positively regulate transcription by recruiting nucleosome remodeling enzymes and histone acetylases (Santos-Rosa et al., 2003; Pray-Grant et al., 2005;). In particular H3K4me1 is established in correspondence of enhancer regions, whereas H3K4me3 to active promoters. When co-present at the same genomic locus, the ratio of H3K4me3 and H3K4me1 signals can suggest the tendency of a particular genomic locus to act as a promoter or enhancers. On the contrary, given that H3K27 acetylation is associated to transcriptional activity, K27 methylation is known to negatively regulate transcription by promoting a compact chromatin structure (Ringrose et al., 2004). A peculiar chromatin modification pattern, consisting of large regions of H3K27 methylation harboring smaller regions of H3K4 methylation, termed “bivalent domains”, was described in mouse ESCs (Berstein et al., 2006). These bivalent domains overlay developmental genes and are thought to maintain them transcribed at very low levels, or completely silent, but poised for activation during cell commitment and differentiation. Indeed, these developmental genes in differentiated cells will be selectively marked by either H3K27 or H3K4 methylation. In this scenario, about two hundreds (N=195) genomic regions were identified as putative neural-commitment enhancers, since epigenetically pre-marked as poised in ESCs (H3K4me1+/H3K27me3+) and active in NECs, loosing H3K27 methylation in favour of acetylation (Rada-Iglesias et al., 2011). These regions were suggested to regulate distal genes associated to gastrulation, germ layer formation, neurulation and early somitogenesis. However, the identification of putative enhancers, and their activity level, by epigenetic signatures could be affected by the particular criteria used for enhancer-definition. Conversely, active promoters and transcriptionally-active enhancers can be directly identified by occurrence of Pol-II transcription and nascent RNAs, which is the direct output of cell regulatory activity and can be revealed by Cap Analysis of Gene Expression (CAGE).

2.7.10Keratinocytes and terminally differentiated keratinocytes

Renewal of epidermis is sustained by stem cells contained in the basal layer and in the bulge of hair follicles. Keratinocyte stem cells (KSCs) have an extensive self-renewal capacity and produce progeny that undergoes terminal differentiation into all the epithelial components of the skin. In this work, long-term keratinocytes were isolated from human neonatal foreskin keratinocytes using a stem cell-enrichment technique based on beta-1

(β1) integrin positive cells, considered as a proxy of KSCs. This method makes use of differences within a keratinocyte culture with distinct differentiation gradients in the ability to adhere to collagen-IV, which is in a direct relationship with the level of β1-integrin expression: cells that adhere more rapidly to collagen-IV express high levels of β1-integrin. The fraction of keratinocytes that adhere within 20 minutes to collagen-IV coated dishes is highly enriched in KSCs (RACs= rapidly adhering cells); in contrast cells that are devoted to terminal differentiation but are still proliferation competent (TA) correspond to the non-adherent cell fraction. The whole culture of keratinocytes in vitro is made up a small percentage of clonogenic cells and a major component of differentiated keratinocytes; the balance between these two groups is shifted to differentiation by culturing keratinocytes for several passages and without their feeder layer support, until exhaustion of their clonogenic potential. Then, the same human keratinocytes used for RAC isolation for different passages on plastic dishes and were obtained a population with a clonogenic fraction < 5-8%, which are considered as the differentiated keratinocyte population (KD) for transcriptome analysis. A comprehensive description of the *epigenome* and *regulome* of KSCs, and of their dynamic changes during commitment and differentiation, is essentially lacking, due to the difficulty in the prospective isolation of a homogenous stem cell population. The identifying of transcriptionally active promoters and regulatory sequences differentially used by human KSCs and their progeny by high-throughput technologies such as CAGE-seq, ChIP-seq is fundamental for a better understanding of the molecular basis of stemness and commitment of human KSCs.

2.8 GSE16256 dataset

GSE16256 derived from UCSD Human Reference Epigenome Mapping Projects (http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/) and contain data ChIP-seq data of H3K4me1, H3K4me3, H3K36me3, H3K9ac, H3K9me3, H3K27me3, H3K4me1, H3K27ac, H2AK5ac, H3K18ac, H3K4me2, H2BK120ac, H2BK12ac, H2BK15ac, H2BK20ac, H2BK5ac, H3K23me2, H3K4ac, H3K56ac, H3K79me1, H3K79me2, H4K20me1, H4K5ac, H4K91ac from H1, H9, IMR90 from H1 cell line, IMR90 cell line, CD34 primary cells, CD34 mobilized primary cells, kidney, fetal day82 F, breast, luminal epithelial cells, breast, myoepithelial cells, breast, stem cells, brain, fetal day122 M, adrenal gland, fetal day96 U, heart, fetal day96 U, kidney, fetal day122 U, lung, fetal day122 U, heart, fetal day101 U, lung, fetal day101 U, CD3 primary cells, CD19 primary cells, ES-WA7 cell line, ES-I3 cell line, CD34 cultured cells, pancreatic islets, iPS-18c cell line, iPS-15b cell line, iPS-20b cell line, iPS-11a cell line, liver, CD15 primary cells, CD4 naive primary cells, breast, vHMEC, peripheral blood mononuclear primary cells, CD8 naive primary cells, lung, fetal day108 F, lung, fetal day103 M, brain, fetal day117 F, lung, fetal day67 M, brain, fetal day85 F, lung, fetal day85 F, brain, fetal day96 F, lung, fetal day96 F, lung, fetal day108 M, H1 +BMP4 cell line, H9 cell line, CD4 memory primary cells, Treg primary cells, adipose nuclei, duodenum mucosa, adipose derived mesenchymal stem cells, bone marrow derived mesenchymal stem cells, iPS-18b cell line, CD4 mobilized primary cells, Th17 primary cells. Libraries were sequenced on different sequencing platforms: Illumina Genome Analyzer (Homo sapiens), Illumina Genome Analyzer II (Homo sapiens), Illumina Genome Analyzer IIx (Homo sapiens), Illumina HiSeq 2000 (Homo sapiens), AB SOLiD 4 System (Homo sapiens), Illumina HiSeq 2500 (Homo sapiens). Details of all protocols can be found at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16256.

For this work, ChIP-seq of H3K4me1, H3K4me3 from H1, H9 cells have been downloaded and analyzed.

2.9 GSE26386 dataset

GSE26320 is composed of two studies from Ernst et al. (2011) and contains ChIP-seq data from 8 distinct chromatin modifications derived from genome-wide histone methylation (H3K4me1, H3K4me2, H3K4me3, H3K27me3, H4K20me, H3K36me3) and acetylation (H3K27ac, H3K9ac) and CTCF maps in embryonic stem cells (H1 ES), erythrocytic leukemia cells (K562), B-lymbhoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblast (NHLF), normal epidermal keratinocytes (NHEK), and mammary epithelial cells (HMEC). Chromatin was fixed and sheared to 200-600 bp using Branson Sonifer. After IP and adapter ligation library fragments of ~250 bp were isolated. The purified DNA was captured on an Illumina flow cell for cluster generation. Libraries were sequenced on Illumina Genome Analyzer II following the manufacturer protocols. The relative gene-expression (Affymetrix) data are available on GSE26312 series. Data for chromatin states are available from ENCODE (http://genome.ucsc.edu/ENCODE) and Gene Expression Omnibus (GSE26312) databases. Data of p300 and DNAseI have been downloaded from ENCODE.

2.10 GSE36994 dataset

GSE36985 refers to Orkin et al. (2012) and contains data from distinct chromatin modification derived from genome-wide histone methylation (H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3) and acetylation (H3K9ac, H3K27ac) maps in adult and fetal CD34+ HSPC-derived proerythroblasts. For ChIP-seq analysis using the HeliScope Single Molecule Sequencer, ChIP DNA was processed for 3' polyA tailing followed by 3' ddATP-blocking. Processing samples by ligation, amplification and size selection are not required by Helicos sequencing. Briefly, 6-9 ng of ChIP DNA or input DNA was used in the 3' polyA tailing reaction in 14.8 ul containing 2 ul of 2.5 mM CoCl2, 2 ul of 10 x terminal transferase buffer and nuclease-free water. The reaction mixture was denatured in 95C for 5 min, followed by rapid cooling in ice water slurry. Then 1 U of terminal transferase, 4 ul of 50 uM dATP and 0.2 ul of NEB BSA were added to the mixture. Samples were incubated in a thermocycler at 37C for 1 h, 70C 10 min, followed by denaturing at 95C for 5 min and rapid cooling in ice water slurry. For 3' ddATP-blocking, 0.5 ul of 200 uM ddATP, 1 ul of 10 x terminal transferase, 1 ul of 2.5 mM CoCl2, 1 U of terminal transferase, and 6.5 ul of nuclear-free water were added to the above reaction. Samples were incubated in a thermocycler at 37C for 1 h and 70C for 20 min. Then 2 picomoles of carrier oligonucleotide were added to the reaction, and samples were hybridized to the Helicos flow cells. In this work I used only data of H3K4me1, H3K4me3, H3K27ac. The relative gene-expression (Affymetrix) data are available on GSE36984 series.

2.11 GSE12646 dataset

GSE12646 describes the experiment of Cui et al. (2009) and contains data from distinct chromatin modification derived from genome-wide histone methylation (H3K4me1, H3K4me3, H3K9me1, H3K9me3, H3K27me1, H3K27me3, H3K36me3, H4K20me1) and

H2AZ, polII maps in CD133+ and CD36+ cells. Sequence reads of mostly 25 bp were obtained using the Solexa Analysis Pipeline. All reads were aligned to the human genome (hg18) and only uniquely mapping reads were retained. In all analyses, reads of multiple identical copies were truncated to contain at most five copies to reduce potential biases. The output of the Solexa Analysis Pipeline was converted to browser extensible data (BED) files detailing the genomic coordinates of each tag. To make the files for visualization on a genome browser, reads originating in non-overlapping 200 bp windows were summed, with the location of reads on positive (negative) strand shifted by +75 bps (-75bps) from its 5' start to represent the center of DNA fragments. The relative gene-expression (Affymetrix) data are available on GSE36984 series.

Chapter 3

Results

3.1 Peak calling

To study the chromatin states and the epigenome of different cell types (hematopoietic, neuronal, endothelial, epithelial) at different differentiation stages (stem cells, progenitors, terminally differentiated cells) I analyzed ChIP-seq data of two chromatin histone marks: H3 lysine 4 trymethylation (H3K4me3), H3 lysine 4 monomethylation (H3Kme1).
Enrichment of H3K4me3 is indicative of active promoters, whereas the presence of H3K4me1 outside promoter regions can be used as a mark for enhancers. ChIP-seq reads were mapped to Hg19 using bowtie, and significant enrichment regions were detected using SICER. SICER is a peak caller algorithm developed to identify regions of signals of histone modifications, that are usually diffuse and lack of well-defined peaks, spanning from several nucleosomes to large domains encompassing multiple genes. For each analysis, the gap-size g parameter, which defines the distance between two significant islands has been chosen as described in Zhang et al. In particular, we examined how the aggregate score of all significant islands change as g is tuned. For chromatin modification with a diffuse profiles e.g. H4K20me1, the aggregate scores increases gradually towards saturation. Then for this type of signal, is suggested to choose a gap size so that the aggregate score is sufficiently close to saturation. Otherwise, in histone modification less broad, the aggregate score quickly reaches maximum, therefore the best g is that maximize the aggregate score. For example, Figure 13 reports the plots of the trends of aggregate score in function of g in 8 histone marks in CD133-cells. Because the high variability of ChIP-seq data (e.g. antibody, number of unique reads mapped to a reference genome) not always this criteria help in the choice of parameters. The best benchmark to analyze ChIP-seq data is therefore tuning the parameter g followed by the visual inspection on a genome browser of the peak calling results.

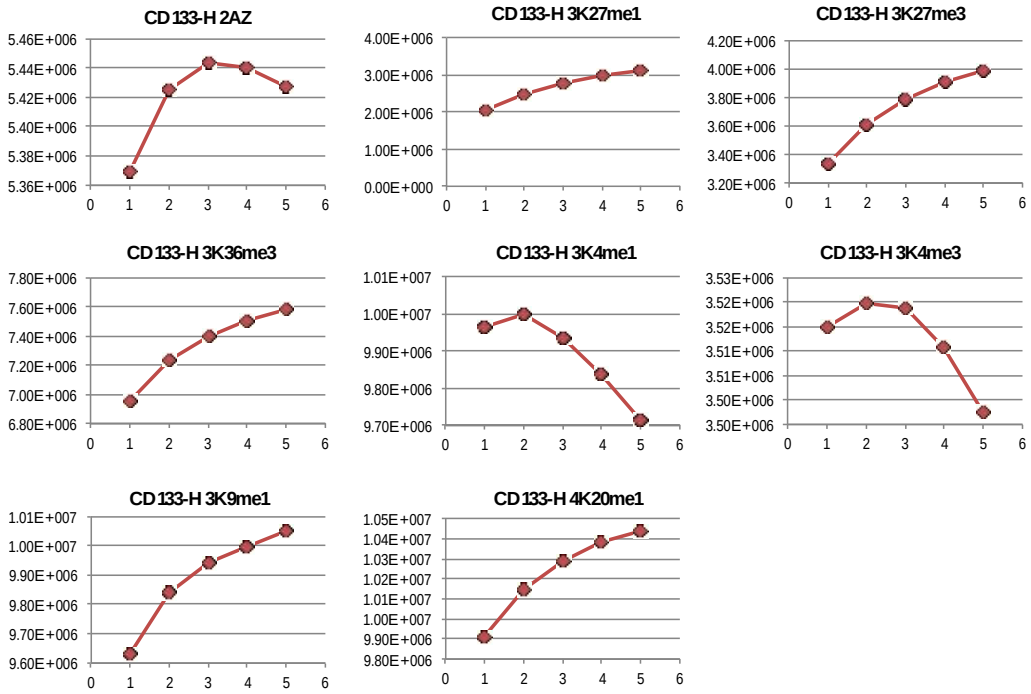

Figure 13: Aggregated score of all significant islands versus gap-size in 8 histone marks from CD133 of Cui et al. experiment.

Using this approach, we identified number of H3K4me1 peaks ranging from 38037 to 115241 in CD36 and H1 cells, respectively, and a number of H3K4me3 peaks ranging from 13890 to 52443 in CD36 and fetal erythroblast. Table 3 reports the number of H3K4me1 and H3K4me3 peaks found in each cell-line.

Table 3. Summary of peak-calling results using SICER

sample	gap-me1	gap-me3	H3K4me1	H3K4me3
CD133	400	400	74400	18930
CD133	1000	600	56397	17713
CD36	800	400	38037	13890
CD34	200	200	103901	52022
CD34	200	200	104892	52443
CD34	400	200	90544	48022
CD34	400	200	91119	52443
Erythroblast_A	400	400	40669	17902
Erythroblast_F	200	400	59695	16883
H1	400	600	115241	18894
H1	400	600	66205	22051
H9	400	600	102893	19379
H9	400	600	57829	21183
KSC	600	400	86435	47500
MYE	400	400	86741	30346
MYE	400	400	114012	38879
NES	400	1000	85488	20918
NES	600	600	70852	22685
NHEK_diff	400	600	105421	32665

3.2 Promotrome construction: statistics and hierarchical structure

CAGE is a technology introduced by Carninci in which the first 20 to 21 nucleotides at the 5' ends of capped mRNA are extracted by a combination of cap trapping and cleavage by restriction enzyme Mme1. There are several advantages that cap-analysis of gene expression offer over standard micro-array approaches to study cell transcriptome. The first is that the expression of all isoforms of a transcript can be inferred. In contrast, the expression measured with micro-array is often a combination of expression of different transcript isoforms. In this work CAGE was used to identify genomic localizations of promoters and to quantify the expression levels of human pluripotent stem cells, myeloblast, erythroblast, embryonic stem cells, neural stem cells, keratinocyte, and differentiated keratinocyte. CAGE data were produced at the Riken laboratory. In this samples, we identified about 16525 TSCs. Within these TSCs there are 3903518 TSSs. we investigated the human promotorome of this cell using structure approach as described in Balwierz et al. (2009). For this aim a custom ad-hoc R script has been develloped. Figure 14 shows the distributions of the number of TSSs per TSC, the number of TSSs per TSR, and the number of TSCs per TSR.

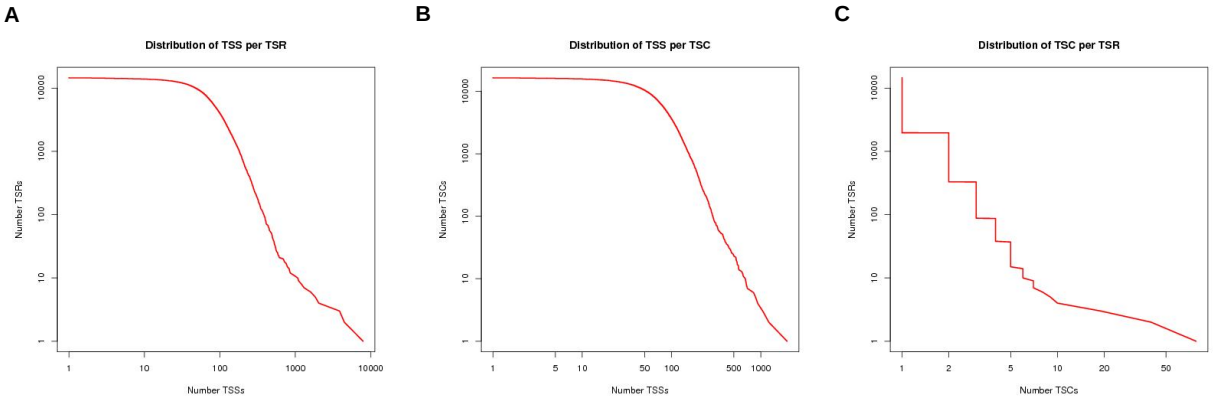

Figure 14: Hierarchical structure of the human promoterome. A. Distribution of the number of TSSs per TSCs. B. Distribution of the number of TSCs per TSR. C. Distribution of the number of TSSs per TSR.

Figure 14C shows that the number of TSCs per TSR is essentially exponentially distributed. That is, it is more common to find only a single TSC per TSR, TSRs with a handful of TSCs are not uncommon, and TSRs with more than ten TSCs are very rare. The number of TSSs per TSC is more widely distribute (Figure 14A). It is more common to find one or two TSSs in a TSC, and the distribution drops quickly with TSS number. However, there is a significant tail of TSCs with between 50 and 150 or so TSSs. The observation that the distribution of the number of TSSs per TSC has two regimes is even clearer from Figure 14A, which shows the distribution of the number of TSSs per TSR. Here again, we see that it is more common to find one or two TSSs per TSR, and that TSRs with between 50 and 150 TSSs are relatively rare. These distributions suggest that there are two different types of promoters. Balwierz and coworkers (2009) suggested that in a *promoterome* exist two class of promotes: *specific promoters* with at most a handful of TSSs in them, and more *fuzzy promoters* with more than ten TSSs. We found different trends with TSCs, *specific promoters* are characterized by a number of TSSs less than of 50, while *fuzzy promoters* TSCs with a

number or TSSs greater than 50.

3.3 CAGE: annotation in HPSCs, eryhtroid and myeloid progenitors

Level 2 promoters were assigned to genes and transcripts (annotated) using public available data sets (RefSeq gene annotated genes, ENSEMBL ncRNA, ncRNA defined by Khalil et al., Cabili et al., and Kretz et al., Vertebrate GenomeAnnotation (Vega) pseudogenes and Yale Gerstein Group pseudogenes). For each set of CAGE level 2 promoters we found, on the same chromosome strand, the gene with the smallest distance to the CAGE-defined level 2 promoters. However, if this distance was greater than 400 bp, the promoters were not annotated. The majority of CAGE level 2 promoters were associated to known genes or transcripts. In particular, annotated promoters accounted for 86% (8305 out of 9675) of all CAGE-defined level 2 promoters in HSPCs (Table 4). Similar frequencies were observed in erythroid and myeloid cells, where 87% (8273 out of 9552) and 87% (8303 out of 9558) of all CAGE-defined level 2 promoters, respectively, were assigned to known genes or transcripts (Table 4). Interestingly, more than 1000 level 2 promoters in each cell population were not associated with any known gene or transcript. These novel promoters may encompass transcription start sites of unknown coding transcripts or ncRNA, such as enhancer RNA (eRNA), recently described by Kim et al. (2010). The number of unique genes and transcripts associated with CAGE level 2 promoters was lower than the number of promoters in the 3 cell types (7759, 7732 and 7761 in HSPCs, erythroid and myeloid progenitors, respectively; Table 4).

Table 4. Annotation of CAGE level 2 promoters. The table shows the number of level 2 promoters, the number of promoters associated with known genes-transcripts, and the number of known genes-transcripts associated with level 2 promoters. Abbreviations: Erythro, erythroid progenitors; Myelo, myeloid progenitors**.**

	# Level 2 promoters	Promoters associated with known genes-transcripts	Promoter NOT associated with known genes-transcripts	Number of known genes-transcripts
HSPCs	9675	8305	1370	7759
Erythro	9552	8273	1279	7732
Myelo	9558	8303	1255	7761

This discrepancy is indicative of alternative promoter usage. In fact, 93.5% of genes-transcripts were associated with 1 CAGE level 2 promoters, whereas transcription of 6.5% of the genes was driven by 2 to 5 alternative promoters (505, 502 and 495 in HSPCs, erythroid and myeloid cells, respectively; Table 5).

Table 5. Distribution of the number of CAGE level 2 promoters per known gene-transcript. The table shows for each hematopoietic cell population the number of CAGE level 2 promoters associated with each known gene-transcript. Abbreviations: *Erythro*, erythroid progenitors; *Myelo*, myeloid progenitors.

# Level 2 promoters	Number of known genes-transcripts		
	HSPCs	*Erythro*	*Myelo*
1	7254	7230	7266
2	468	467	453
3	34	32	38
4	2	2	3
5	1	1	1
Total	7759	7732	7761

In some cases transcripts were generated by the usage of alternative promoters in a lineage-specific fashion. For example, we observed that *LMO2* gene, coding for a transcription factor essential hematopoietic development, is transcribed from 3 different promoters (Promoter 1, 2 and 3; Figure 15). Promoter 1 was used only by HSPCs at low levels, instead Promoter 3 was mainly used by HSPCs and myeloid cells, whereas *LMO2*-associated transcripts in erythroid cells arose mostly from Promoter 2 (Figure 15).

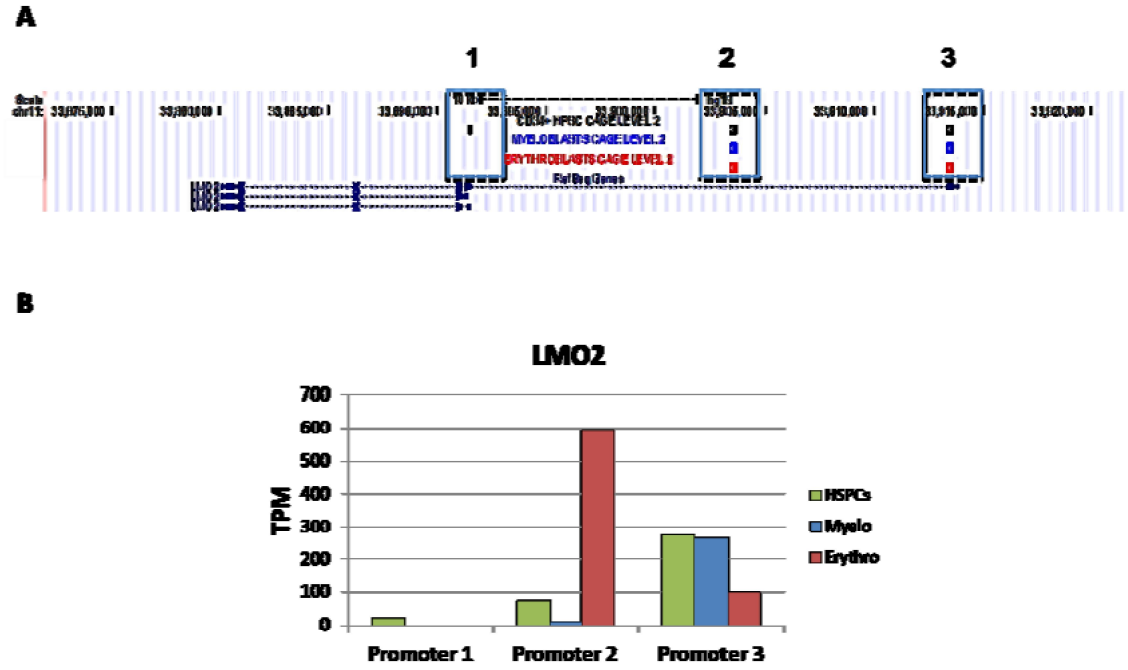

Figure 15. *LMO2* alternative promoters. (A) The 3 different promoters of *LMO2* gene are highlighted with blue boxes. (B). The expression levels of *LMO2* promoters in each hematopoietic cell population are shown. Abbreviations: Erythro, erythroid progenitors; Myelo, myeloid progenitors.

Then, we analyzed the proportion of CAGE level 2 promoters, associated to different classes of RNA, i.e., coding RNA and ncRNA (lincRNA, miRNA, rRNA, snoRNA and snRNA) (Figure 16). In both HSPCs and their lineage-committed progeny, most of CAGE level 2 promoters were associated with coding RNAs (75% in HSPC and 76% in myeloid and erythroid cells). 10 to 11% of the promoters were assigned to ncRNAs, whereas 13 to 14% were not associated with any transcript.

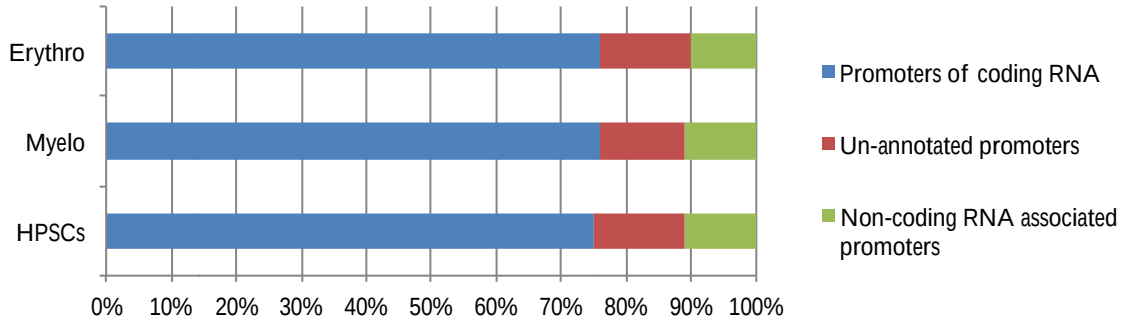

Figure 16. Distribution of CAGE level 2 promoters. The graph show the proportions of CAGE level 2 promoters associated with annotated promoters, un-annotated promoters, and non-coding of RNA (i.e., coding RNA, ncRNA, and lincRNA).

3.4 CAGE: annotation in hES and NSC

CAGE level 2 promoters were annotated on the basis of their vicinity to RefSeq genes, in order to associated each sequenced promoter with its gene or transcripts. The majority of mapped CAGE clusters/promoters was found in an interval of 400 bp, from the nearest Refseq sequence, and was assigned as promoters of that gene; in hES Refseq associated promoter accounted for 75% (10.828) of all CAGE clusters, a results similar to NSC, where 70% (7392) of all CAGE clusters were assigned to Refseq genes (Figure 17). The remaining non-RefSeq CAGE clusters (2419 for hES and 3076 for NSC) were considered as novel promoters of unknown transcripts and were mostly found in exonic, intronic, and intergenic positions. The ~ 3% of cage level2 promoters were annotated in correspondence of ncRNA both in hES and NSC cells.

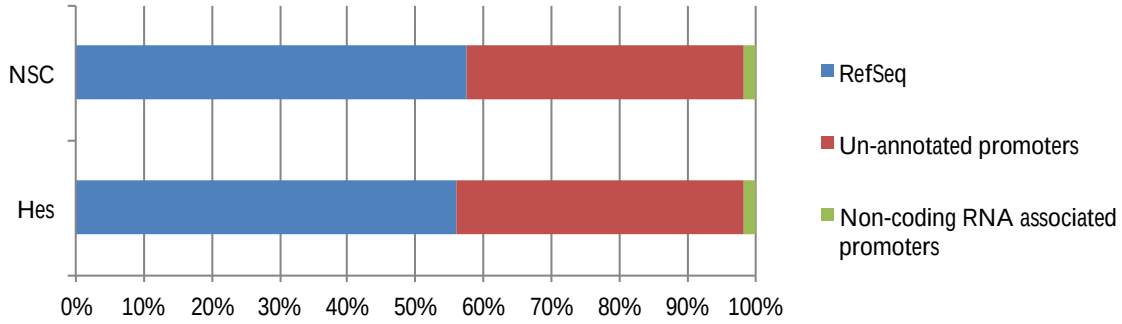

Figure 17. Proportions of CAGE level 2 promoters associated with annotated promoters, un-annotated promoters, and non-coding of RNA (i.e., coding RNA, ncRNA, and lincRNA).

3.5 CAGE: annotation in keratinocyte stem cells (KSC) and differentiated keratinocytes (NHEK)

The same analysis of annotation was performed on keratinocyte stem cells (KSC/RAC) and differentiated kernatinocyte (KD). CAGE clusters/promoters were annotated in an interval of 400 bp from the nearest known transcript sequence (Refseq) and were assigned as promoters of that gene; in RACs Refseq associated promoter accounted for 74.3% (10.828) of all CAGE clusters, a results similar to differentiation keratinocyte, where 73.1% (10.985) of all CAGE clusters were assigned to Refseq genes (Table 6 and Figure 18). Finally, the remaining CAGE clusters (3737 for RACs and 4042 for KDs) were considered as novel promoters of unknown transcripts and were mostly found in exonic, intronic, and intergenic positions. Finally the ~ 25% and 26 % of cage level2 promoters were annotated in correspondence of ncRNAs in RAC and KD cells respectively.

Table 6. Deep-CAGE sequencing of promoters engaged in RAC and KD populations. The table resumes the total number of promoter actively used and mapped by CAGE-seq in RAC and KD, the fraction of promoters associated to a RefSeq gene, and the number of promoters not associated to a RefSeq gene or associated to a ncRNA.

	Expressed promoters	Promoters associated with RefSeq genes	Promoters associated with ncRNAs.
RAC	14565	10828	3737
KD	15027	10985	4042

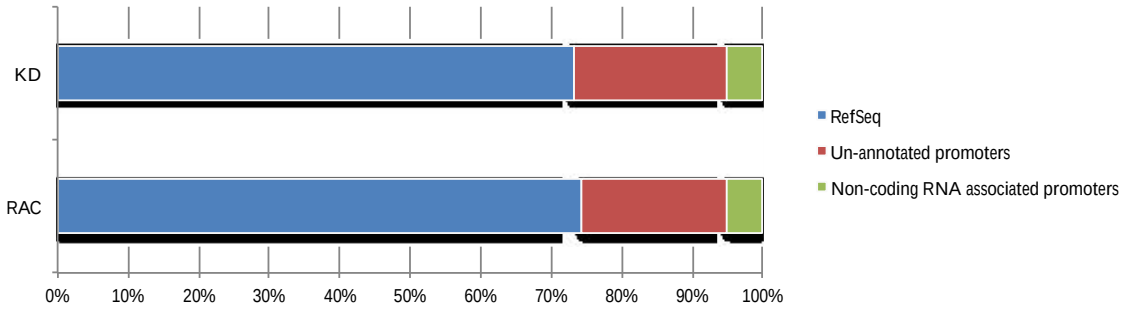

Figure 18. Proportions of CAGE level 2 promoters associated with annotated promoters, un-annotated promoters, and non-coding of RNA (i.e., coding RNA, ncRNA, and lincRNA).

3.6 Detection of promoter and enhancers using IRCM

The identification of cis-regulatory modules (CRMs) is of pivotal importance to clarify the, still poorly understood, molecular mechanisms underlying fundamental characteristics of human stem cells, such as self-renewal, commitment, and differentiation. A deeper knowledge of these mechanisms is crucial not only to shed light on basic cellular mechanisms but also to envision stem cell-based therapies. The differentiation of cells into distinct tissues and organs requires the expression of specific sets of genes at each developmental stage and in each cell type. A major challenge in biology is to comprehensively and accurately identifying the DNA sequences that are required to regulate gene expression, i.e., *cis-regulatory* elements or *cis-regulatory modules*. Recently, ChIP-seq has been developed to identify cis-regulatory elements in the genomes of several organisms including Humans Sapiens, Drosophila Melanogaster and Caenorhabditis elegant. Here, to detect enhancers and promoters using two histone marks and CAGE data, we considered ChIP-seq profiles of H3K4me1 and H3K4me3 from different cellular types, i.e. embryonic, blood, and keratinocyte. For each of these categories, we used cells at different developmental stages. The data set was build using data from public repositories or obtained at the Center for Regenerative Medicine of the University of Modena and Reggio Emilia (CMR; Table1). Raw sequences of every histone mark have been aligned to reference 19 of the human genome using bowtie (Langmead et al., 2009). Peak calling in ChIP-seq data has been performed using SICER (Zang et al., 2009). CAGE data for human pluripotent stem cells, myeloblast, erythroblast, embryonic stem cells, neural stem cells, keratinocyte, and differentiated keratinocyte were produced at the Riken laboratory. In the last years, different bioinformatics tools have been designed to analyze ChIP-seq data and to characterize genome regulatory elements, as Hidden Markov Models (Ernst et al., 2011), dynamic Bayesian networks (Hoffman et al., 2012), and profile methods (Heintzman et al., 2007). Despite all these efforts, no standard method has been developed to detect cis-regulatory yet, mostly due to different experimental design (numbers of analyzed histone marks) and to a lack of consensus on the rules to detect cis-regulatory modules. In fact, while it is well accepted that active-promoters co-localize in regions with low levels of H3K4me1 and high levels of H3K4me3, it is not clear which histone-marks or proteins co-localize with enhancers. The problem is further complicated by the fact that enhancers are tissue-specific. In general high levels of H3K4me1 and low levels of H3K4me3 characterize strong-enhancers, but other histone marks like H3K27ac, H3K4me3, and H3K9ac may be informative to detect enhancers. In general if only H3K4me1 or H3K4me3 composes the experimental model, the characterization of promoters and

enhancers can be performed using a simple peak-calling algorithm. This method is adequate in the discovery of promoters because they are well characterized by enrichment of H3K4me3. In contrast, the genome-wide localization of enhancers using only significant regions of H3K4me1 is advised and produce a roughly map of enhancers in the genome. When the experimental model is composed both by H3K4me1 and H3K4me3 the identification of promoters and enhancers can be realized using computational methods that codify biology patterns. In general, high levels of H3K4me3 and low levels of H3K4me1 characterize promoters. Then an algorithm able to identify this chromatin state is appropriate to identify cis-regulatory elements. In Barksi et al. (2007), the authors developed a method that using H3K4me1 and H3K4me3 regions estimate the ratio of reads between them in correspondence of polII sites. The sites with high levels of H3K4me3 are defined such *putative promoters*, otherwise *putative enhancers*. When H3K4me1, H3K4me3 and p300 compose the experimental model, the most appropriate algorithm to use is the one designed by Heintzmann and collaborators based on profile matching. Finally if the number of histone marks is greater than 3 for each cell-lines, the use of bioinformatics methodologies based on unsupervised and supervised machine learning methods is the choice suggested. In this work, I developed a custom R-workflow to identify promoters and enhancers using two histone marks. This is method based on intersect-ratio-count-method (IRCM), an approach similar to those used in Barksi et al. This computational pipeline allows elucidating the global *regulome* in deeper and detailed manner. In particular IRCM allows extracting three categories of cis-regulatory elements:

i) those characterized by the co-localization of H3K4me1 and H3K4me3, H3K4me1+/H3K4me3+);
ii) the peaks of H3K4me1 and H3K4me3 that do not overlap between them on the genome H3K4me1+/H3K4me3-//H3K4me1-/H3K4me3+;
iii) the islands of H3K4me1 or H3K4me3 that overlap with cap analysis of gene expression data, H3K4me1+/CAGE-seq// H3K4me3+/CAGE-seq.

Briefly, the pipeline comprises three steps and uses as input the files with the significant islands generated by SICER. In the first step, the script invokes BEDtools (Quinlan et al., 2010) to identify regions where H3K4me1 peaks overlap (do not overlap) with H3K4me3. The second step considers only the overlapping regions, for which log-ratios between of H3K4me3 and H3K4me1 tag counts are quantified. The tag counts of H3K4me3 and H3K4me1 are first normalized using the sequencing depths of both libraries. Then, if the ratio is greater than 0, the region is labeled as putative promoter, otherwise as putative enhancer. Finally, CAGE regions are merged with ChIP-seq peaks using ChIPpeakAnno (Zhu et al., 2010), considering a distance of 2000 kb between a ChIP-seq and CAGE peak. Considering only the numbers peaks of H3K4me3 using SICER, we identified from 13890 to 52443 putative promoters region in CD36 and CD34 cells (see the table 3 for the details of the other cells). With IRCM, the numbers of putative enhancers regions characterized by the presence H3K4me1+/H3K4me3+ is of 7400 enhancers in CD36 cells and 25181 in neural derived stell cells (H1-der). Using the same combination of chromatin state we recognized 4344, 31774 promoters respectively in CD133 and H1 cells (Figure 19). IRCM recovered all promoters and enhancers marks by the only presence on the genome of H3K4me3+ or H3K4me1+ (Figure 20). Using these criteria 43789 and 87484 enhancers were discovered in CD133-cells and hES (H9). Instead, 602 and 38640 peaks of H3K4me3 that does not overlap with H3K4me1 were localized in ERY and H1 cells, respectively (for the complete list of these putative-enhancers/promoters regions refer to Table 20). Considering the H4K4me1/CAGE-seq we found a range of 712, 1943 enhancers,

respectively in fetal erythroblast and CD133 cells. As to promoters, we characterized 60 H3K4me3+/CAGE-seq regions in CD34 and 6147 in H9. Intriguing the number of CAGE peaks that overlap with epigenetic marks, in general, is greater in correspondence of H3K4me3 peaks, than H3K4me1 (Figure 21). This is consistent with the capability of cap analysis of gene expression to identify the localization of each transcription start site. Finally, to further assess the ability of the pipeline to predict cis-regulatory elements, we run the same analysis using public data of adult and fetal ProESHuman primary adult pro-erythroblasts (Xu et al., 2012). The original analysis identified 10790 and 10687 active promoters (characterized by the only presence of H3K4me3) in fetal ProES and adult ProES, respectively. As a result, we found 8393 promoters in adult ProES and 8054 in fetal ProES that correspond to the ~78% and ~75% of promoters found in the original publication.

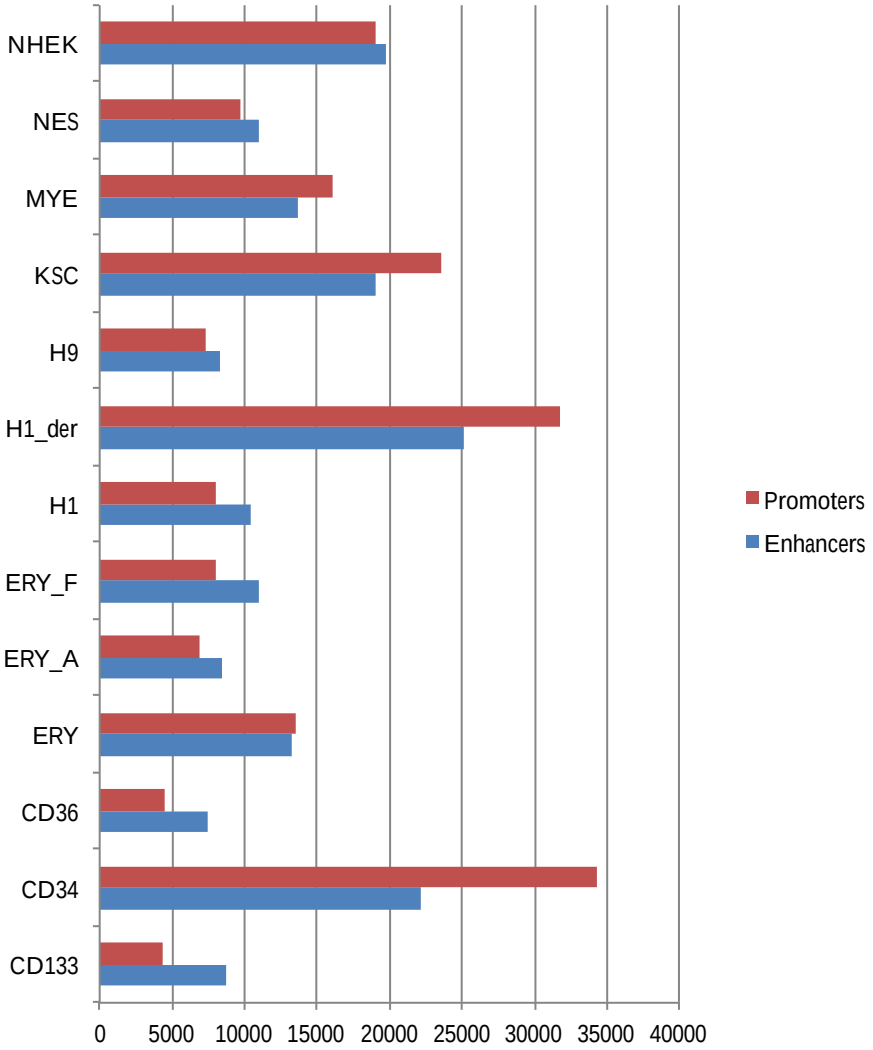

Figure 19. Number of promoters and enhancers characterized using both H3K4me1 and H3K4me3 (H3K4me1+/H3K4me3+)

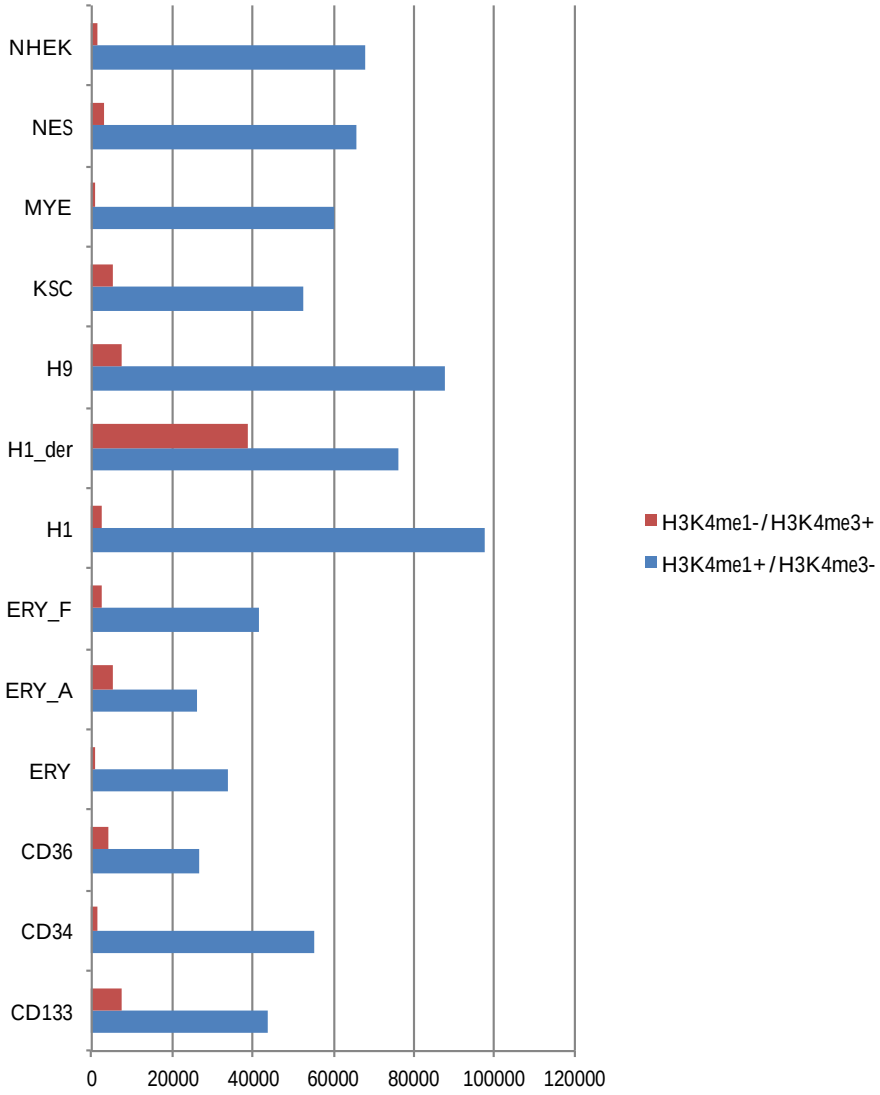

Figure 20. Number of promoters and enhancers by H3K4me1+/H3K4me3-, H3K4me1-/H3K4me3+

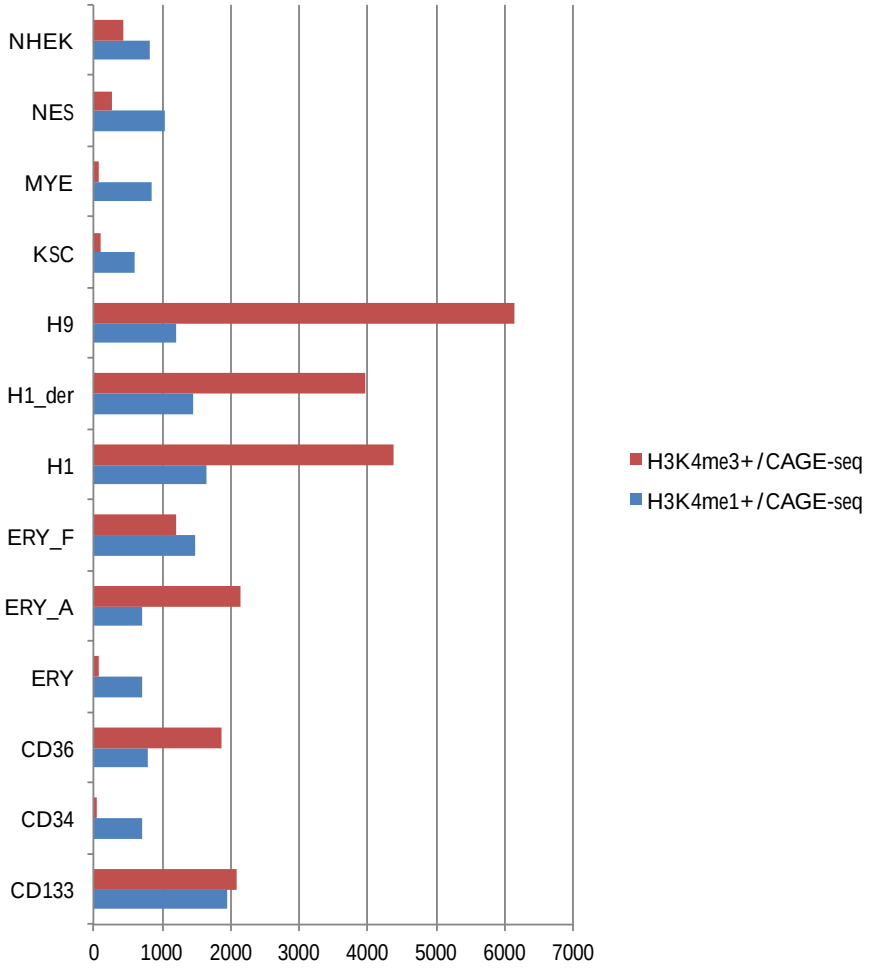

Figure 21. Number of promoters and enhancers characterized by H3K4me1+/CAGE-seq//H3K4me3+/CAGE-seq

In the human genome, enhancers are associated with active chromatin marks, in a cell-type-specific manner, in contrast to promoters that are ubiquitously occupied in multiple cell lines. Here we used the entropy of Shannon to verify the tissue-specificity usage of promoters and enhancers H3K4me1+/H3K4me3+ of embryonic, blood, and keratinocyte at different stage of development cells. For this aim, we developed a custom R script based on *entropy* package. In each system of differentiation cell we found that the occupancy of enhancers H3K4me1+/H3K4me3+ is tissue-specific. In contrast, promoters characterized by the same chromatin pattern are used in manner ubiquitous during the differentiation in each cell line. These results are consistent with other outcomes described in Shen et al (2012) and in Barrera et al (2008). Figures 22, 23, and 24 report the tissue-specificity index for each category of cis-regulatory elements, measured by the entropy score. An entropy close to zero indicates the occupancy at this sites is highly tissue-specific, while an entropy score close to $\log_2(N)$ means the sites binds uniformly.

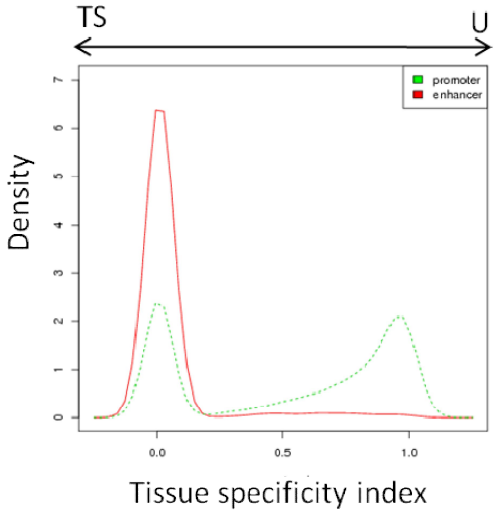

Figure 22. Tissue specificity of the usage of promoters (H3K4me1+/H3K4me3+) and enhancers (H3K4me1+/H3K4me3+) in CD34+/CD133+, MYE, ERY

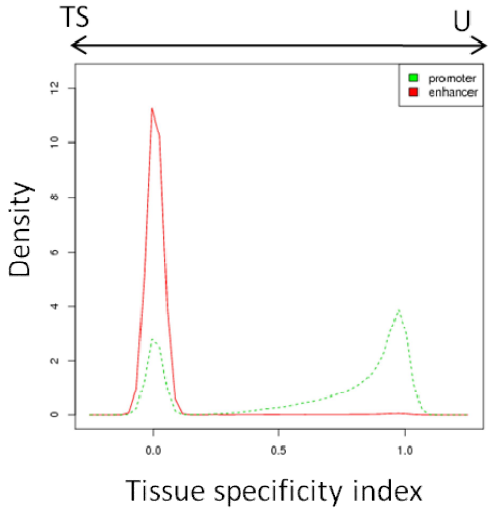

Figure 23. Tissue specificity of the usage of promoters (H3K4me1+/H3K4me3+) and enhancers (H3K4me1+/H3K4me3+) in hES (H9) and NES

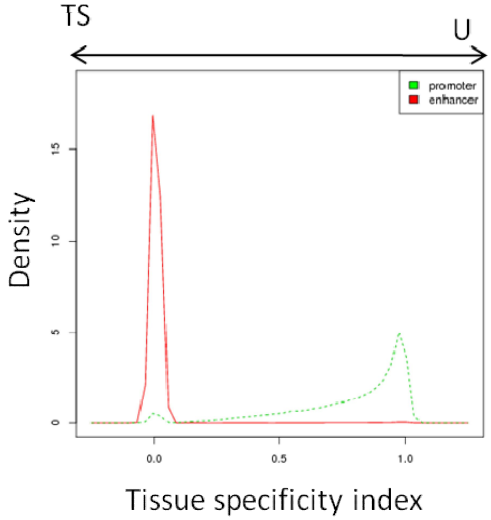

Figure 24: Tissue specificity of the usage of promoters (H3K4me1+/H3K4me3+) and enhancers (H3K4me1+/H3K4me3+) in KSC, NHEK

Therefore we examined mono and try methylation of histone H3 lysine 4 (H3K4me3, H3K4me1) at the promoters (H3K4me3+/H3K4me1+) found using IRCM reasoning that the state of these histone modifications would vary in a cell-type-specific manner. Figure 25 shows the enrichments of reads in H3K4me3 and H3K4me1 at promoters and enhancers H3K4me1+/H3K4me3+ during the differentiation of CD133 in erythroid and myeloid progenitors. The chromatin signatures at promoters are remarkably similar across all cell types. Whereas the chromatin patterns at predicted enhancers (H3Kme3+/H3Kme1+) are distribute in a cell-type-specific manner. To evaluate the dynamics of putative promoters and enhancers' chromatin signatures upon hematopoietic stem cell differentiation, we compared H3K4me3 and H3K4me1 islands in HSPCs and erythroid and myeloid lineages, finding that 14870 H3K4me3 islands and 28023 H3K4me1 islands were shared by the 3 populations (Figure 25).

In hES and NSC cells, promoters show highly conservative patterns. In contrast, the enhancers have little rate of conservation chromatin signatures (Figure 26). During the differentiation of keratinocyte stem cells into terminated differentiated cells, distal regulatory elements shown different patterns at chromatin states (Figure 27). These results are concordant with the analysis of tissues specificity using Shannon entropy and with the results described in Heintzman et al. 2009. The heat maps were created using SeqMiner (Ye T et al., 2010), and clustered with kmeans using 10 clusters. To further investigate the usage of the cis-regulatory elements I studied the dynamics of promoters and enhancers during differentiation in each cellular model. Most of HSPCs H3K4me3 islands overlapped with erythroid and myeloid promoters (92% and 90%, respectively), whereas a lower proportion of HSPCs H3K4me1 islands are shared with erythroid and myeloid progenitors (61% and 47%, respectively; Figure 28). These results suggested that the same HSPCs putative promoters are used in lineage-restricted progenitors, while HSPCs putative enhancers consistently change upon erythroid and myeloid differentiation. While the 85% of hES H3K4me3 regions overlapped with NES putative promoters. In contrast the 50% of enhancers are shared between hES and NES. During the differentiation of KSC in differentiated keratynocyte the 43% and 40% of promoters and enhancers are shared between them respectively. Therefore, it is likely that the fine regulation of commitment and differentiation, which involve small changes the transcriptional profile in HPSC, kSC, KSC, is carried out by regulatory elements (putative enhancers) that are differentially used in different stages of lineage commitment (Figure 28).

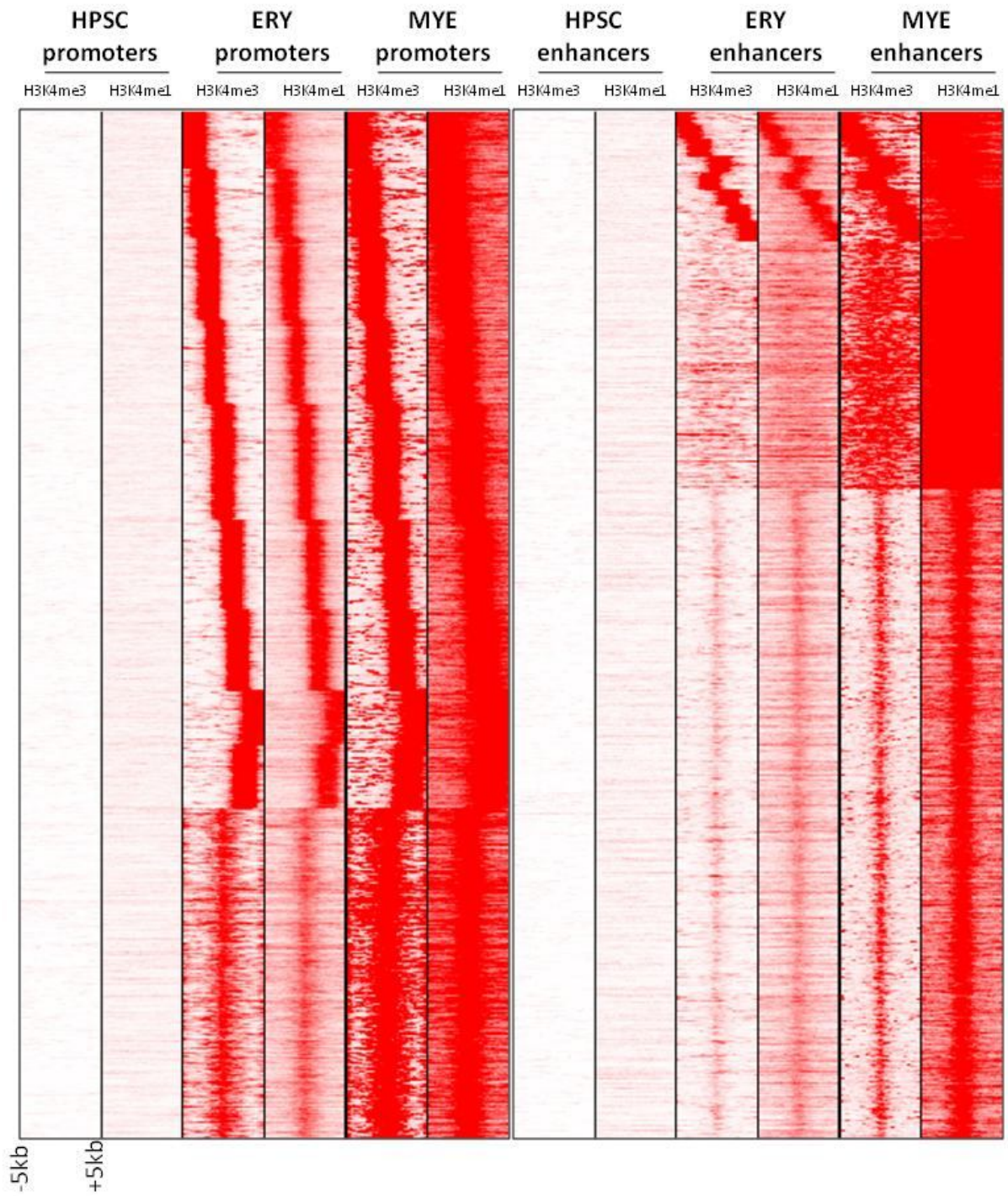

Figure 25. Chromatin modifications at promoters are generally cell-type invariant whereas those at enhancers are cell-type-specific. ChIP-seq data have been used to map histone modification (H3K4me1, H3K4me3) in two cell types (HPSC, ERY, MYE). The heat map was generated using seqminer ± 5 kb from 12270 non redundant promoters, and observe them to be generally invariant across cell- types. In clustering on 9410 non-redundant enhancers predicted with IRCM on the basis of chromatin marks, revealing the cell-type specificity of enhancers.

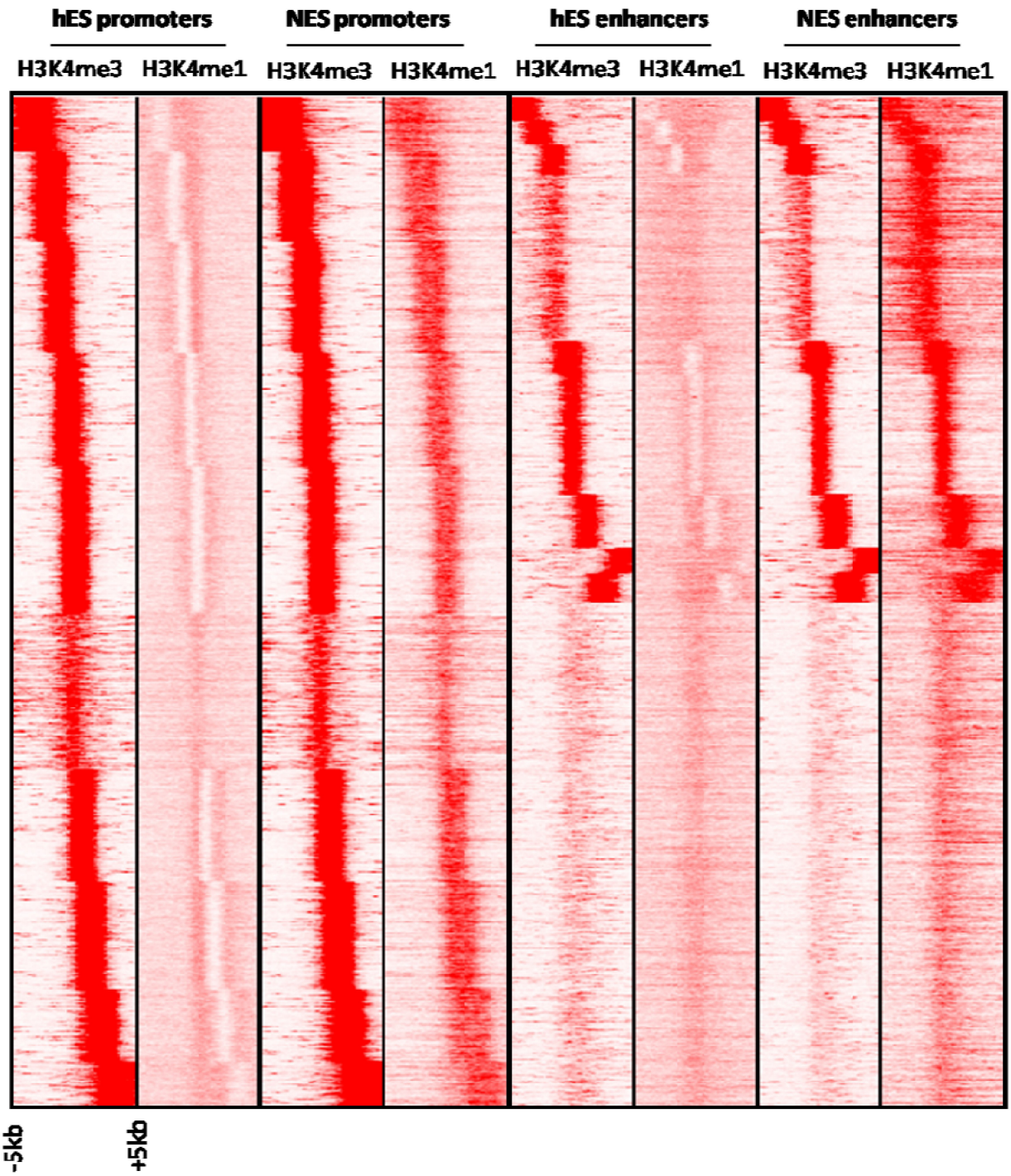

Figure 26. Chromatin modifications at promoters are generally cell-type invariant whereas those at enhancers are cell-type-specific. ChIP-seq data were used to map histone modification (H3K4me1, H3K4me3) in two cell types (hES, NES). The heat map was generated using seqminer ± 5 kb from 12270 non redundant promoters, and observe them to be generally invariant across cell- types. In clustering on 9410 non-redundant enhancers predicted with IRCM on the basis of chromatin marks, revealing the cell-type specificity of enhancers.

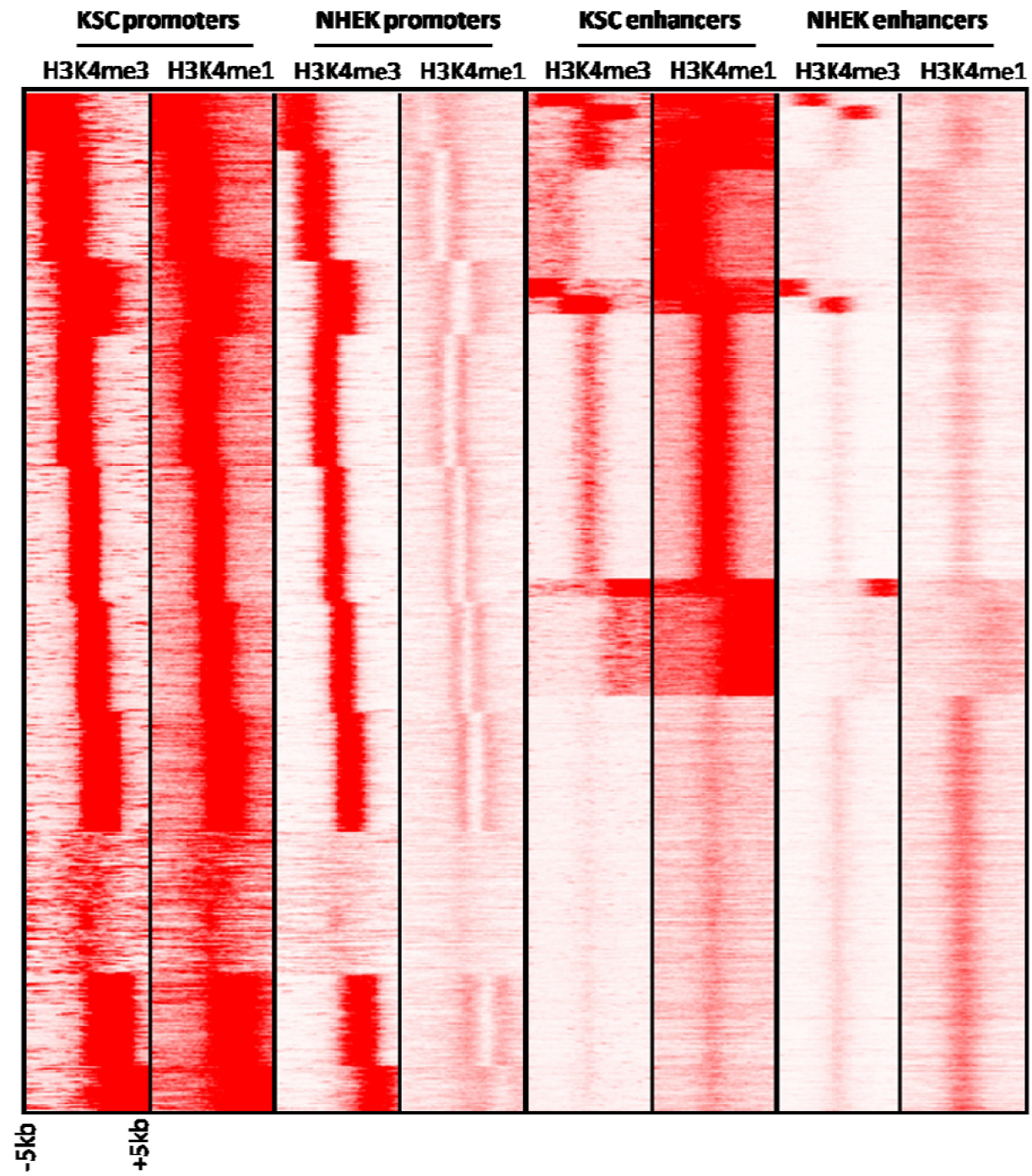

Figure 27: Chromatin modifications at promoters are generally cell-type invariant whereas those at enhancers are cell-type-specific. ChIP-seq data was used to map histone modification (H3K4me1, H3K4me3) in two cell types (KSC, NHEK). The heat map was generated using seqminer ±5 kb from 27734 non redundant promoters, and observe them to be generally invariant across cell- types. In clustering on 21209 non-redundant enhancers predicted with IRCM on the basis of chromatin marks, revealing the cell-type specificity of enhancers.

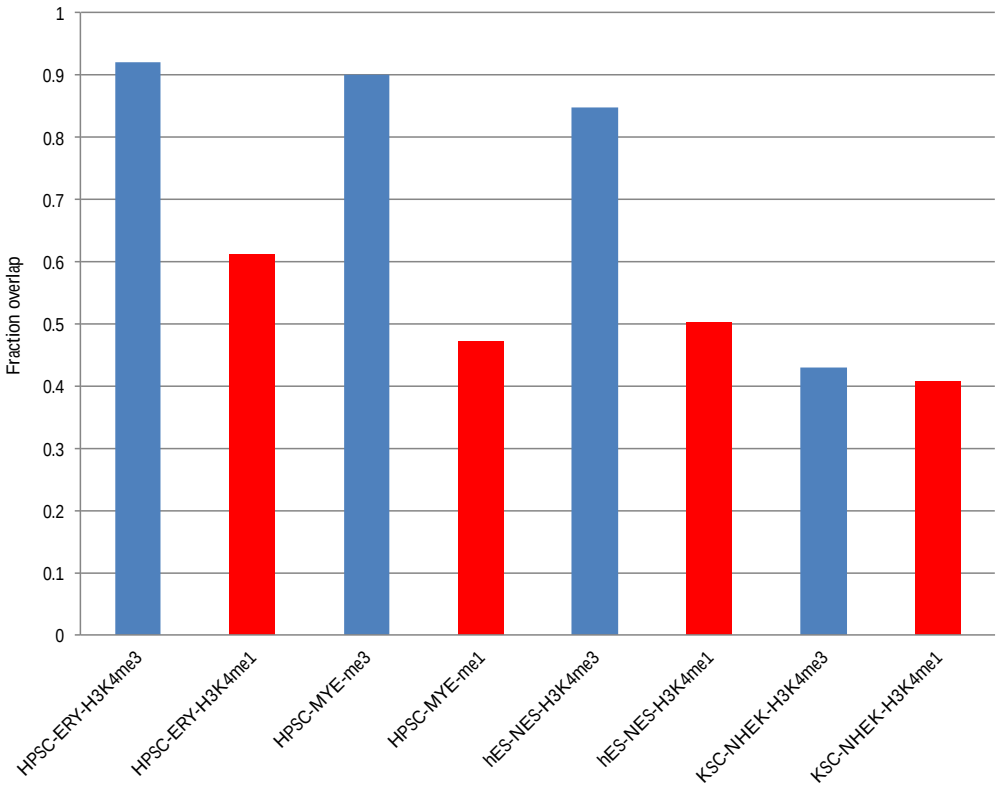

Figure 28. Dynamics of promoter and enhancer chromatin signatures during the differentiation of HSPC, hES, KSC. The plot shows the proportions of HSPC H3K4me3 and H3K4me1 islands shared between stamina and differentiated cells.

3.7 Prediction of promoters and enhancers using SVM2CRM

The regulation of a gene is an exceedingly complex process controller by interacting *proximal* and distal DNA sequence elements usually placed in a *cis* configuration. The *proximal* element is the basal promoter where the general transcription machinery assembles. Promoters are always located in close proximity to the 5'-end of a gene. *Enhancers* are distal elements that bind transcriptional factors and activate the transcription of the target gene. Enhancers have been studied using experimental techniques such electrophoretic mobility shift assay (EMSA), molecular cloning with a reporter gene and mutation analyses. During the years, several studies have shown that enhancers share sequence conservation with orthologous regions in other mammalian genomes. Therefore, different bioinformatics tools were developed to predict enhancers using the sequence conservation (Aparicio et al., 1995). However the use of this approach have some weaknesses. Firstly, that this analysis does not distinguish active or poised states of the chromatin. Secondly, is missing the information about under condition (e.g. stage of cell development) a particular enhancer is active. Finally, because this algorithms use the sequence conservation to predict enhancers, all not-conserved enhancers are not considered. The experimental investigation of the proteins and chemical modifications associated with enhancers and promoters, especially by ChIP-chip, and later by ChIP-seq showed that the post-translational modification of the histones, including methylation, acetylation is linked to specific events, including transcriptional activation, heterochromatin formation, chromosomal segregation. The acetylation and methylation of specific histones

is of particular interest in the study of gene regulation as these effects determine the activation, inactivation, of *cis-regulatory elements.* As described in the above sections, enhancers are involves in cell development and differentiation of the many different cell types in the body by activating cell type-specific gene expression programs. This is clearly illustrated in hematopoiesis, where distinct gene expression programs characterize various hematopoietic progenitors. Recently, with the discovery of different domains of histones that specific to enhancers, has encouraged the genome-wide identification of enhancers. This approach involves the identification of chromatin methylations, by ChIP-seq, followed by the application of pattern recognition algorithms trained with specific chromatin profile signatures. In these methods, aligned tag counts are processed into profiles of specified windows sizes with positive profiles, centered at enhancer marker-enrichment regions, background profiles are created at random regions. Then, classifiers are implemented to discriminate cis-regulatory elements from background profiles. As mentioned above, these methods are limited as they either consider only a small number of epigenetic marks. The profile method (PM) and Hidden Markov model method (HMM) only explored datasets using a limited number of chromatin modifications. While, artificial (Firpi et al., 2010) method CSI-ANN combined ~40 dataset of methylation and acetylation signatures. Here I developed, SVM2CRM a novel method to uncover chromatin states. SVM2CRM implements *liblinear* and *multicore,* respectively two libraries for classification of large data-set, and a package to run the analysis in parallel. The results of ChIP-seq analysis produce huge dataset characterized also by the presence of long numeric vector with value 0. Therefore, we choose *liblinear*, because is able to reduce the computational cost of predictions and to manage big dataset with millions of features as ChIP-seq data. Moreover, *liblinear* use *sparseM*, another package to manage sparse matrices as ChIP-seq data. This is useful, because avoid to perform some preprocessing step of the ChIP-seq signals. For e.g. in some cases is necessary download from UCSC tables of black regions, these are characterized by the absence of ChIP-seq signals and computationally are long vectors with values 0. Independently from the methodologies used to perform classification, this *sparse* data can have an influence and to introduce biases during the predictions, therefore during the preprocessing of ChIP-seq data is suggested to remove these regions. Because *liblinear* implements *sparseM,* the preprocessing step just described is not necessary. SVM2CRM contain 26 functions essential for preprocessing of input, preprocessing of data before of prediction, tuning of parameters (e.g. kernel, cost function, number of histone marks; Table 7). Moreover, functions to plot the results of tuning and of pre-processing have been designed. As described in the Material and Methods chapter, SVM2CRM consists of four steps: i) data preprocessing ii) tuning of SVM parameters iii) predictions of regulatory regions iv) output generation.

Here we tested SVM2CRM to recognize enhancers from the ChIP-seq signals of 8 genome-wide histone methylation and acetylation maps obtained from stem cells (H1 ES), erythrocytic leukemia cells (K562), B-lymbhoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblast (NHLF), normal epidermal keratinocytes (NHEK), and mammary epithelial cells (HMEC). We designed, trained, and tested a series of SVM models using increasing numbers of histone marks, ranging from 2 (e.g., H3K4me1, H3K4me3) to 8 marks (H3K4me1, H3K4me3, H3K27ac, H3K4me2, H3K27me3, H3K9ac, H3K36me3, H4K20me1) in each cell-line. Using the *tuningCombinationParameters* function we generated in total 355056 models. Moreover, other 349992 models were generated using *cisREfind* function with shape parameter, that defines

how to SVM2CRM model the signals of the histone marks around true positive features such p300 binding sites. In total we generated 662898 models and used different criteria to estimate the performance of predictions.

Table 7. Lists of functions implemented in SVM2CRM

Function	step
applyKmeans.R	*preprocessing-input*
bestSVMHMcomb.R	*tuning parameters*
bestSVMLinearHMcomb.R	*tuning parameters*
buildTrainingSet.R	*preprocessing-prediction*
cisREfindPM.R	*preprocessing-prediction*
cisREfind.R	*preprocessing-prediction*
clusteringProfileKmeans.R	*preprocessing-prediction*
createBed.R	*write output*
createModel.R	*preprocessing-prediction*
createSVMinput.R	*preprocessing-prediction*
fastSW.R	*preprocessing-prediction*
findFeatureOverlap.R	*preprocessing-input*
gridFilter.R	*tuning parameters*
medianClusterResults.R	*preprocessing-input*
mergeModel.R	*preprocessing-input*
parseOutputfastSW.R	*preprocessing-input*
performanceSVM.R	*preprocessing-input*
plotBestKcluster.R	*plot*
plotEstimatePerformance.R	*plot*
plotTuningModel.R	*plot*
searchBestTrainingSet.R	*tuning parameters*
takeRegion.R	*preprocessing-input*
tuningParametersComb.R	*tuning parameters*
tuningParameters.R	*tuning parameters*
weigthHMmodel.R	*tuning parameters*

As described in Material and Methods, *performanceSVM* allows computing *sensitivity*, *specificity*, *fscore*. To assess the performance of predictions we focused on *fscore* as described in Fernandez et al. (2012). Then we also tested the effects on the performance of prediction of different SVM kernel types L2-regularized logistic regression (k=0), L1-regularized logistic regression (k=6), L2-regularized logistic regression-dual (k=7), distributions of examples (w), and cost functions. Usually, in the definition of the weights between two classes to predict enhancers, the suggested correct ratio is 1:10. Here the positive classes are defined using the position of p300 binding sites, while the negative classes are randomly selected from background regions of the genome. In each cell line considering the default kernel (i.e., L2-regularized logistic regression-dual) the *fscore* is higher when models are trained with a distribution of enhancers/random regions of 1:7. Figure 29-30. Moreover, the kernel type does not seem to significantly affect the prediction performances. For e.g. in GM12878, the values of *fscore* of the models predicted using k=0, L2 regularized logistic regression, K=7 L2 regularized logistic regression (dual) are similar, little difference in the values of *fscore* are present in K=6 L1-regularized logistic regression Figure 31). This can be explain because K=0, L2 regularized logistic regression and K=7 L2 regularized logistic regression dual are based on the same algorithm but the computation of the results is computationally different. Therefore the *L2 regularized logistic regression* kernel, that is the default parameter used in *LiblineaR* function could be use as parameter to predict enhancers.

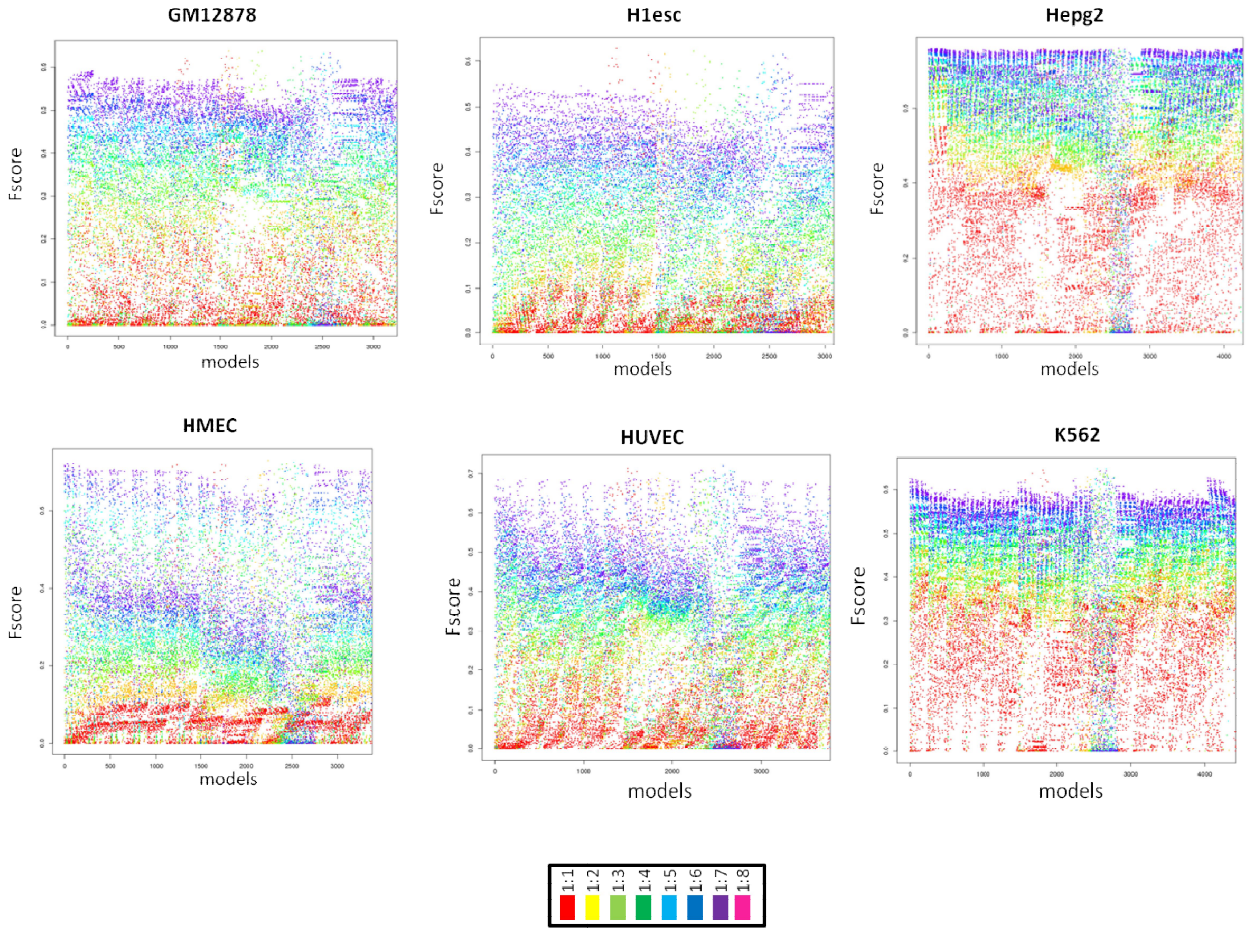

Figure 29. The influence of weights between the two classes (enhancers/not enhancers) on the fscore in Ernst dataset (B-lymbhoblastoid cells (GM12878), stem cells (H), hepatocellular carcinoma cells (HepG2), mammary epithelial cells (HMEC), umbilical vein endothelial cells (HUVEC), erythrocytic leukemia cells (K562)). On x-axis the number of models, on y-axis the fscore

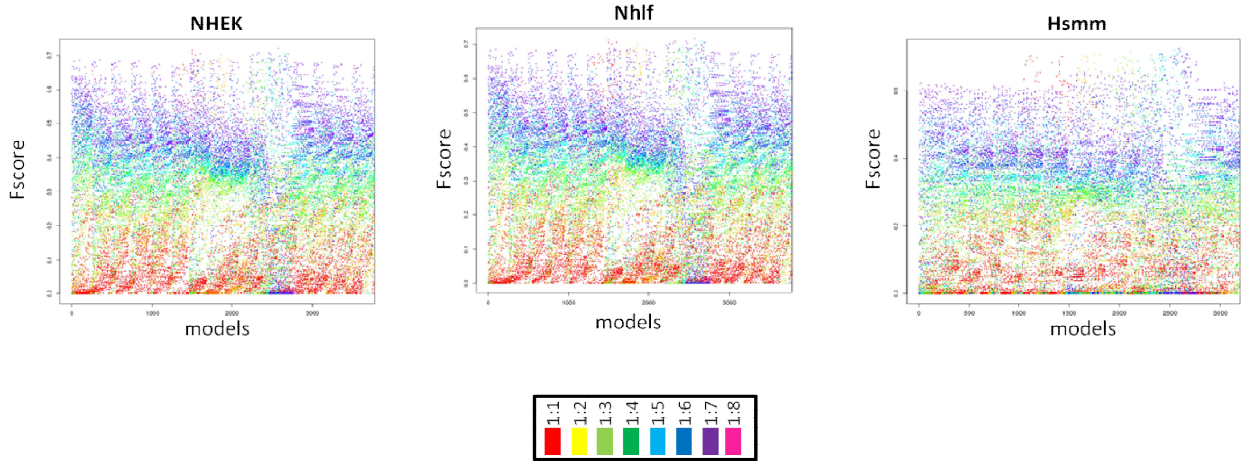

Figure 30. The influence of weights between the two classes (enhancers/not enhancers) on the fscore in Ernst dataset (normal epidermal keratinocytes (NHEK), normal lung fibroblast (NHLF), skeletal muscle myoblasts (HSMM)). On x-axis the number of models, on y-axis the fscore

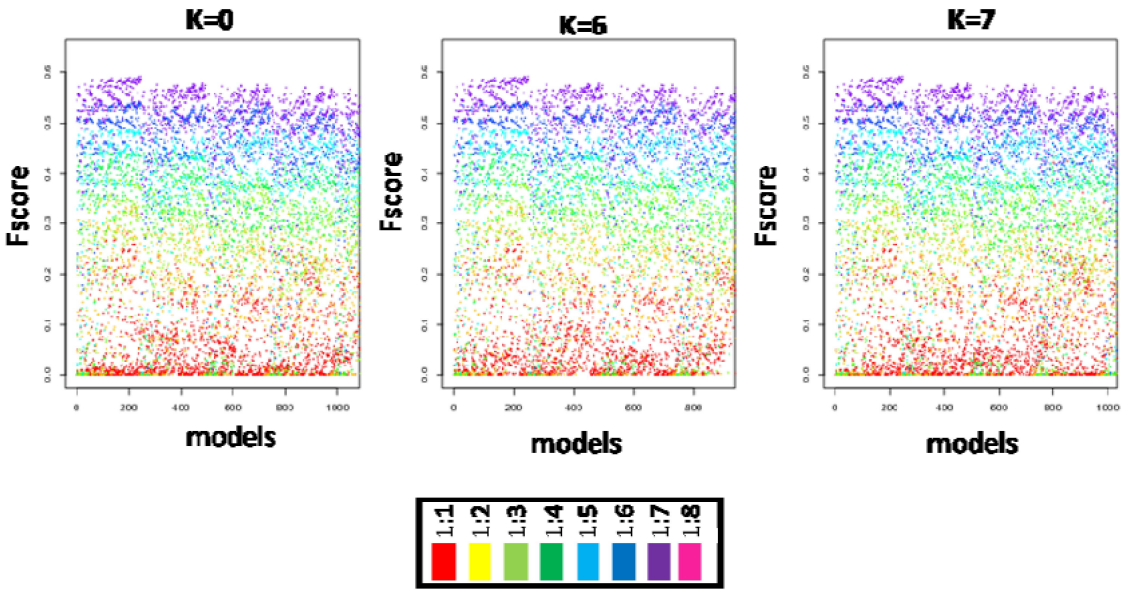

Figure 31. The influence of kernel and weights between the two classes (enhancers/not enhancers) on the fscore in GM12878. On x-axis the number of models, on y-axis the fscore

Considering ChIP-seq data of GM12878 cell line, SVM2CRM created 37057 possible combinations of models (different k, cost, weights of classes). Selecting the 1071 models built with k=7 and w=1:7, the top performing model on a 10-fold cross-validation (fscore= 0.65) resulted the one using H3K4me1, H3K4me2, H3K27ac, and H3K9ac (Figure 32). Using this model, SVM2CRM detected 13044 putative-enhancers. This results are consistent with the biology: mono and tri-methylation of histone H3 lysine 4 (H3K4me1, H3K4me2) are known signatures of enhancers, while acetylation of histone H3 lysine 9 and 27 define active transcriptional states. Therefore the chromatin states H3K4me1+/ H3K4me2+/ H3K27ac+/ H3K9ac+ characterize active enhancers. Then I performed a comparative analysis of the performance of SVM2CRM and three other predictive approaches on the same ENCODE region of GM12878 (Table 8). A large fraction of predicted promoters overlapped with experimental evidences. We predicted a total of 13044 enhancers versus 5605, 682 reported by PM and CSIANN. Then, we computed the percentage of predicted enhancers that overlap with p300 and DHS sites. The selected model recovered the 19% and 22.53% of p300 and DHS sites, respectively. SVM2CRM predicted 13044 enhancers in GM12878 of which ~73.7% overlapped with at least one type of experimental evidence (table 8, Figure 33). Finally, we compared the performance of SVM2CRM with others published methods, as CSI-ANN and profile-matching, finding that SVM2CRM presented the highest precision in GM12878, outperforming CSI-ANN and PM methods by ~3% and ~21%, respectively (table 9). These evidences suggest that SVM2CRM is able to discovery enhancers from histone methylation and acetylation maps.

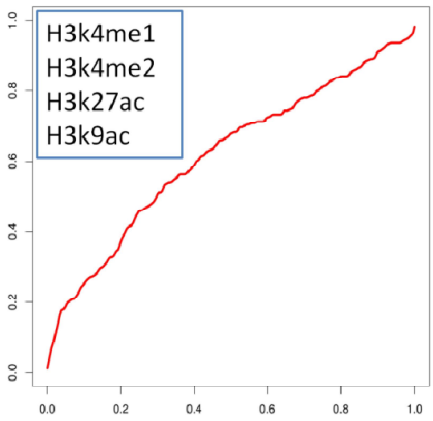

Figure 32. Cross validation ROC plot of the optimum SVM model to predict enhancers in GM12878 using H3K4me1+/ H3K4me2+/ H3K27ac+/ H3K9ac+

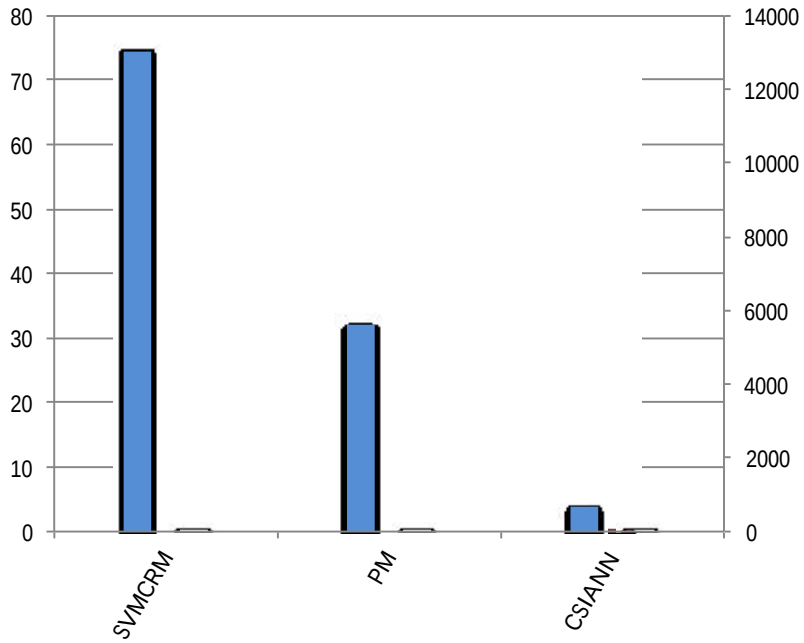

Figure 33. Number of enhancers predicted using SVM2CRM, PM, CSIANN that overlaps with p300 and DHS. The red bar represent the percentage of p300+dhs that overlaps with the total of predicted enhancers (blue bars)

Table 8 Total number of true experimental evidence of p300 and DHS, #overlap of p300 and DHS, % of overlap between predicted promoters and true experimental evidence

Algorithm	**p300 (tot)**	**# overlap p300**	**p300(%)**
SVM2CRM	13497	2518	18.65
PM	13497	863	6.39
CSIANN	13497	30	0.22
Algorithm	**DHS (tot)**	**# overlap DHS**	**DHS(%)**
SVM2CRM	31494	7097	22.53
PM	31494	2021	6.41
CSIANN	31494	12	0.03

Table 9. Number of predicted enhancer in SVM2CRM, PM, CSIANN, percentage of overlap with true experimental evidence, and precision

Algorithm	# predicted enhancer	TP (p300)	precision (%)
SVM2CRM	13044	2518	19.30
PM	5605	863	15.39
CSIANN	682	30	4.39
Algorithm	**# predicted enhancer**	**TP (DHS)**	**precision (%)**
SVM2CRM	13044	7097	54.40
PM	5605	2021	36.05
CSIANN	682	12	1.76

Finally, SVM2CRM was used to predicted enhancers in H1esc, Hmec, Hpeg, Hsmm, HUVEC, K562, NHEK, Nhlf. From the 349992 models we selected only those tuned with w=1:7 because is the best parameter to optimize the fscore. Using this first step of filtering, for each cell line I reduced the numbers of models to ~4899. Then from remaining tables we selected all models with fscore greater than 0.5 and false positive rate <0.2. Using this approach we identified the best models for each cell-line with the best combinations of histone marks, weight W, kernel (K), cost C to perform the classification. The values of fscore in the different cell-type ranged from 0.50 in H1esc to 0.75 in Hmec. The number of enhancers predicted in chromosome 1 and 2 is included from 19430 of NHEK cell to 40262 of HUVEC (Table 10). Of these enhancers, the 40% and 85.65% overlaps with p300 binding sites in H1esc and Huvec, while the 34.48% and the 38.09% overlaps with DHS sites in H1esc and Hpeg (Table 11).

Table 10. Chromatin histone marks found in the best models, W (weights between two classes), K=kernel, C=cost, #enhancer, total number of predicted enhancers, fscore, sensitivity, precision

Sample	Histone marks	W	K	C	Predicted enhancers	fscore	Sens	Prec
H1esc	H3K4me1//H3k4me2//H3k36me3//H4k20me1//H3k9ac	1:7	0	0.001	27998	0.50	0.33	0.90
Hmec	H3K4me1//H3k4me2//H3k27me3//H4k20me1	1:7	0	0.001	21800	0.71	0.60	0.87
Hpeg	H3K4me1//H3k4me2//H3k27ac//H3k36me3	1:7	6	0.01	22376	0.75	0.67	0.84
Hsmm	H3K4me1//H3k4me2//H3k36me3//H4k20me1	1:7	6	10	33115	0.59	0.47	0.82
HUVEC	H3K4me1//H3k36me3	1:7	3	100	40262	0.68	0.59	0.79
K562	H3K4me1//H3k4me2	1:7	0	0.001	25412	0.60	0.47	0.84
NHEK	H3K4me1//H3k4me2//H4k20me1	1:7	6	0.01	19430	0.65	0.51	0.89
Nhlf	H3K4me1//H3K4me3//H4k20me1	1:7	6	10	20281	0.66	0.54	0.87

Table 11. Chromatin histone marks found in the best models: overlap with p300 and DHS sites

Sample	#Tot (p300)	#p300	#p300%	#Tot -DHS	#DHS	#DHS%	#P300+DHS
H1esc	1854	1588	85.65	43238	14909	34.48124	16497
Hmec				58718	20466	34.85473	20466
Hpeg	4644	4092	88.11	36763	14006	38.09809	18098
Hsmm				54284	19045	35.084	19045
HUVEC				33384	12202	36.55044	12202
K562	1919	783	40.80	38228	13507	35.33274	14290
NHEK				38871	14203	36.53881	14203
Nhlf				37073	13561	36.57918	13561

Then by the visual inspection we verified the genome localization of predicted enhancers using integrate genome browser (Figure 34). The predicted enhancers fall in

correspondence of the gene bodies or intergenic regions. As described in the Material and Methods section, enhancers are considered as distal regulatory elements; therefore their presence far from TSSs regions is consistent.
Therefore, we observed a concordance in the genomic coordinates between predicted enhancers and true sites of binding of p300 or with distal regulatory elements predicted with analysis of chromatin states (Ernst et al., 2011).

Figure 34. From the top of the picture: in red the true binding sites of p300, in blue the predicted enhancers, in orange the chromatin states

SVM2CRM discovered enhancers from histone modification data recognizing epigenetics pattern associated with p300-enrichment regions. Unfortunately, p300 binds only a reduced subset of enhancers therefore the total number of real enhancers far to be known. Since non-functional regions (true negatives or false positives) are difficult to define, the experimentally and or computationally supported regions should be as part of the real enhancer repertoire. This evidences suggest that the best quality metrics to assess the performance of prediction is the precision rather than by its sensitivity. Because when the positive examples available constitute a very reduced subset of the positive space, it is crucial to avoid over-fitting of the training data. Using ChIP-seq data of GM128678 we found that the precision of SVM2CRM is the highest among all the methods under analysis. Moreover, in others cell lines the percentage of overlap between predicted enhancers and true positive sites is greater than 34.48%. SVM2CRM is a robust implementation that over previously published methods in predicting enhancers.

Chapter 4

Conclusions

The human body is composed of diverse cell types with distinct functions. Although it is known that lineage specification depends on cell-specific gene expression, which in turn is driven by promoters, enhancers, insulator and other cis-regulatory DNA elements for each gene, the relative roles of these regulatory elements in this process are not clear. Among these, enhancers play a central role in driving cell-type-specific gene expression and are capable of activating transcription of their target genes at great distances ranging from several to hundreds, in rare cases even thousands, of kilobases. High throughput sequencing technologies, combined with the knowledge that cis-regulatory elements are associated with certain chromatin features, can be effectively used to annotate enhancers and promoters. A large number of efforts aimed at genome-wide enhancers annotation in a variety of cell types and organisms, the most extensive of which is represented by the ENCODE Project. These studies not only confirmed that enhancers are the most dynamically utilized part of the genome, but also revealed a staggering number of putative enhancers elements, with 400.000 distinct enhancers mapped in a defined set of human cell lines, and current estimate of cis-regulatory elements harbored by the human genome at over a million. During these years several computational algorithms were designed to identify cis-regulatory elements. These bioinformatics tools are based on biological features of enhancers: the sequence conservation (Aparicio et al., 1995), the presence transcription factor binding sites (Chens et al., 2012), and the localization with specific patterns of chromatin states (Heintzman et al., 2007). All this computational methods have some disadvantages. For instance, methods based on sequence conservation are only able to detect enhancers that are orthologous between species. Moreover with the increase of ChIP-seq data and new uncovered in the epigenomic research of the last years, tools based on the discovery of chromatin patterns (e.g. profile matching) are becoming inadequate. At last, despite all these efforts, cis-regulatory maps are available only for few cell-lines.

This thesis addressed two main problems related with the i) development of new bioinformatics algorithms to predict functional elements in the genome and ii) the discovery of promoters and enhancers in several human cell-types for which cis-regulatory maps are not available. Specifically, we analyzed ChIP-seq and Cap Analysis of Gene Expression (CAGE) data from 7 cell-lines, i.e., HPSC, erythroid and myeloid progenitors and embryonic stem cells and neural stem cells, keratinocyte stem cells, and differentiated

keratinocytes, and considered antibodies for histone H3 lysine 4 try-methylation (H3K4me3), a modification associated with promoters, and for H3 lysine 4 mono-methylation (H3K4me1), preferentially associated with enhancers. Moreover, public datasets were used to complete the panel of histone modifications for HPSC and hES (H9). Finally we tested a R package to predict enhancers on a dataset of ChIP-seq data that consists of embryonic stem cells (H1 ES), erythrocytic leukaemia cells (K562), B-lymphoblastoid cells (GM12878), hepatocellular carcinoma cells (HepG2), umbilical vein endothelial cells (HUVEC), skeletal muscle myoblasts (HSMM), normal lung fibroblasts (NHLF), normal epidermal keratinocytes (NHEK) and mammary epithelial cells (HMEC). In this case, we considered antibodies for histone H3 lysine 4 trimethylation (H3K4me3), a modification associated with promoters; H3K4me2 (dimethylation), associated with promoters and enhancers; H3K4me1 (methylation), preferentially associated with enhancers; lysine 9 acetylation (H3K9ac) and H3K27ac, associated with active regulatory regions; H3K36me3 and H4K20me1, associated with transcribed regions; H3K27me3, associated with Polycomb-repressed regions.

All these analysis required an integrative approach both to merge ChIP-seq data between them or with CAGE peaks. In the dataset that contains ChIP-seq and CAGE data from hematopoietic, neuronal, endothelial, epithelial we tested different methodologies for data integration and complexity reduction. The complexity reduction method was used to identify the number of significant enrichment regions for H3K4me1 and H3K4me3 and to give a rough estimate of the number of putative enhancers and promoters. Then the outputs from these analyses were used to investigate rigorously the number of promoters and enhancers using the Intersect Reads Count Method (IRCM), a novel method that we developed to identify cis-regulatory elements.

Finally, we developed a R package (named *SVM2CRM*) that allows the detection of cis-regulatory elements using machine-learning methods and an implementation of profile matching (Heintzman et al., 2007). SVM2CRM was implemented using *liblinear* and *multicore,* two R packages that perform classification of two classes on big matrises using parallel computing. This method circumvents some limitations of previously available bioinformatics tools since it integrates two different models to predict enhancers. The first one, based on profile methods is useful when the dataset of chromatin marks comprises H34K4me1, H3K4me3 and p300 or any others transcription factors or regions with a list of true sites of enhancers. The second allows performing classification when the numbers of antibodies used to map particular histone marks is greater than three. Also in this case it is mandatory the presence of a list of true regions known to be enhancers. Given a set of histone marks, SVM2CRM estimates the best combinations of support vector machine parameters and chromatin marks to predict enhancers. We tested SVM2CRM using public available ChIP-seq data, then we performed a comparative analysis of the performance predictions between SVM2CRM, profile methods, and CSI-ANN. SVM2CRM was used with increasing numbers of histone marks, ranging from 2 (e.g., H3K4me1, H3K4me3) to 8 (H3K4me1, H3K4me3, H3K27ac, H3K4me2, H3K27me3, H3K9ac, H3K36me3, H4K20me1). Using this approach we elucidated that L2-regularized logistic and L2-regularized logistic (dual) kernels guarantee similar prediction performances.

Then, we reduced the numbers of models from 37057 to 1071 using L2-regularized dual kernel (k=7) and weights between classes of 1:7, the top performing model on a 10-fold cross-validation (fscore= 0.65) resulted the one using H3K4me1, H3k4me2, H3k27ac, and H3k9ac. Using this model, SVM2CRM detected 13044 putative-enhancers. Considering the percentage of predicted enhancers that overlap with p300 and DHS sites, the selected

model recovered the 19% and 22.53% of p300 and DHS sites, respectively. SVM2CRM predicted 13044 enhancers in GM12878 of which ~73.7% overlapped with at least one type of experimental evidence. Finally, we compared the performance of SVM2CRM with others predictors, as CSI-ANN and profile-matching, finding that SVM2CRM presented the highest precision in the given dataset, outperforming CSI-ANN and profile-matching by ~3% and ~21%, respectively. Using the same procedure, we mapped enhancers in H1, HPEG2, Hmec, Hsmm, Nhek, Nhlf. Globally we identified 210674 putative enhancers: of these, the 85% overlaps with p300 binding sites in H1 and 40.80% in K562, while the 34.48% and 38.09% overlaps with DHS sites in Hpeg and H1 cells.

In conclusion, the research activity in this thesis aimed at developing novel bioinformatics algorithms for the genome-wide prediction of cis-regulatory elements. The availability of computational pipelines to identify functional elements is of pivotal importance to elucidate biological processes underlying self-renewal, differentiation and commitment of cells and therefore the development of organs and tissues.

Chapter 5

References

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.

Aparicio S, A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby et al. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes Proc Natl Acad Sci U S A, 92 (1995), pp. 1684–1688

Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. 2009;10(7):R79. doi: 10.1186/gb-2009-10-7-r79.

Barrera, L.O. et al. Genomewide mapping and analysis of active promoters in mouse embryonic stem cells and adult organs. Genome Res 18, 46□59 (2008).

B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal marginclassifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pages 144{152. ACM Press, 1992.

Berstein LM, Tsyrlina EV, Poroshina TE, Levina VV, Vasilyev DA, Kovalenko IG, Semiglazov VF. Genotoxic factors associated with the development ofreceptor-negative breast cancer: potential role of the phenomenon of switching of estrogen effects. Exp Oncol. 2006 Mar;28(1):64-9.

Bhatia, M., Wang, J. C., Kapp, U., Bonnet, D. & Dick, J. E. Purification of primitive human hematopoietic cells capable of repopulating immune-deficient mice. Proceedings of the National Academy of Sciences of the United States of America94, 5320–5 (1997).

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory, pages 144–152, New York, NY, 1992.

Cabili, M. N. et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes & development25, 1915–27 (2011).

Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrst AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M,Konno H, Nakano K, Mottagui-Taber S, Arner P, Chesi A, Gustincich S, Persichetti F, et al.: Genome-wide analysis ofmammalian promoter architecture and evolution. Nat Genet 2006, 38:626-635. (FANTOM3)

C. Cortes and V. N. Vapnik. Support vector networks. Machine Learning, 20:1–25, 1995.

C.Y. Chen, Q. Morris, J. Mitchell ,Enhancer identification in mouse embryonic stem cells using integrative modeling of chromatin and genomic features

Cui K, Zang C, Roh TY, Schones DE et al. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. Cell Stem Cell 2009 Jan 9;4(1):80-93.

Doulatov, S., Notta, F., Laurenti, E. & Dick, J. E. Hematopoiesis: a human perspective. Cell stem cell10, 120–36 (2012).

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, Ku M, Durham T, Kellis M, Bernstein BE. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011 May 5;473(7345):43-9

Faulkner, G. J. et al. A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics91, 281–8 (2008).

Fernández M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines.

Firpi ,H.A., Ucar,D. and Tan,K. (2010) Discover regulatory DNA elements using chromatin signatures and artificial neural network. Bioinformatics, 26, 1579–1586.

Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J,Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R. GENCODE: producing a reference annotation for ENCODE. Genome Biol.2006;7 Suppl 1:S4.1-9.

Hausser, J., and K. Strimmer. 2009. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. J. Mach. Learn. Res. 10: 1469-1484.

Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet. 2007 Mar;39(3):311-8

Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012 Mar 18;9(5):473-6

Khalil, A. M. et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proceedings of the National Academy of Sciences of the United States of America106, 11667–72 (2009)

Kharchenko PK, Tolstorukov MY, Park PJ. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol. 2008 Dec;26(12):1351-9

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. Widespread transcription at neuronal activity-regulated enhancers. Nature. 2010 May 13;465(7295):182-7

Kretz M, Webster DE, Flockhart RJ, Lee CS, Zehnder A, Lopez-Pajares V, Qu K, Zheng GX, Chow J, Kim GE, Rinn JL, Chang HY, Siprashvili Z, Khavari PA. Suppression of

progenitor differentiation requires the long noncoding RNA ANCR. Genes Dev. 2012 Feb 15;26(4):338-43.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25

Lassmann T, Frings O, Sonnhammer EL: Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. Nucleic Acids Res 2009, 37:858-865

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet. 2010 Jul;11(7):476-86

Elkabetz Y, Panagiotakos G, Al Shamy G, Socci ND, Tabar V, Studer L. Human ES cell-derived neural rosettes reveal a functionally distinct early neural stem cell stage. Genes Dev. 2008 Jan 15;22(2):152-65.

Li et al. (2008) SOAP: short oligonucleotide alignment program". BIOINFORMATICS, 24 no.5,713–714

McLean Y Cory, Dave Bristor, Michael Hiller, Shoa L Clarke, Bruce T Schaar, Craig B Lowe, Aaron M Wenger, and Gill Bejerano. "GREAT improves functional interpretation of cis-regulatory regions". Nat. Biotechnol. 28(5):495-501, 2010. PMID 20436461

McPherson JD et al., A physical map of the human genome. Nature. 2001 Feb 15;409(6822):934-41.

Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. Science, 302, 413.

Orkin, S. H. Diversification of hematopoietic stem cells to specific lineages. Nature reviews. Genetics1, 57–64 (2000).

Patterson PH, Nawa H. Neuronal differentiation factors/cytokines and synaptic plasticity.

Cell. 1993 Jan;72 Suppl:123-37

Pennacchio,L.A., Ahituv,N., Moses,A.M., Prabhakar,S., Nobrega,M.A., Shoukry,M., Minovitsky,S., Dubchak,I., Holt,A., Lewis,K.D. et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. Nature, 444, 499–502.

Pray-Grant MG, Daniel JA, Schieltz D, Yates JR 3rd, Grant PA. Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature. 2005 Jan 27;433(7024):434-8.

Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2

Rada-Iglesias A, Wysocka J. Epigenomics of human embryonic stem cells and induced pluripotent stem cells: insights into pluripotency and implications for disease. Genome Med. 2011 Jun 7;3(6):36.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics.2010 Jan 1;26(1):139-40.

Romualdi, C., Bortoluzzi, S. & Danieli, G. a Detecting differentially expressed genes in multiple tag sampling experiments: comparative evaluation of statistical tests. Human molecular genetics10, 2133–41 (2001).

Ringrose L, Paro R. Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins. Annu Rev Genet. 2004;38:413-43.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and **C.-J. Lin**. LIBLINEAR: A library for large linear classification Journal of Machine Learning Research 9(2008), 1871-1874.

Santos-Rosa H, Valls E, Kouzarides T, Martínez-Balbás M. Mechanisms of P/CAF auto-acetylation. Nucleic Acids Res. 2003 Aug 1;31(15):4285-92.

Shen Y, Yue F, McCleary DF, Ye Z,Edsall L, Kuan S, Wagner U, DixonJ, Lee L, Lobanenkov VV, Ren B. A map of the cis-regulatory sequences in the mouse genome. Nature. 2012 Aug 2;488(7409):116-20.

Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y: Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci USA 2003, 100:15776-15781.

Tropepe V, Hitoshi S, Sirard C, Mak TW, Rossant J, van der Kooy D. Direct neural fate specification from embryonic stem cells: a primitive mammalian neural stem cell stage acquired through a default mechanism. Neuron. 2001 Apr;30(1):65-78.

Urbanek S. (2009). Multicore: parallel processing of R code on machines with multiple cores or CPUs. R package version 0.1-4. http://www.rforge.net/multicore/.

Visel,A., Rubin,E.M. and Pennacchio,L.A. (2009) Genomic views of distant-acting enhancers. Nature, 461, 199–205.

Wilming, L. G. et al. The vertebrate genome annotation (Vega) database. Nucleic acids research36, D753–60 (2008).

Wang, J. C., Doedens, M. & Dick, J. E. Primitive human hematopoietic cells are enriched in cord blood compared with adult bone marrow or mobilized peripheral blood as measured by the quantitative in vivo SCID-repopulating cell assay. Blood89, 3919–24 (1997).

Xu J, Shao Z, Glass K, Bauer DE, Pinello L, Van Handel B, Hou S, Stamatoyannopoulos JA, Mikkola HK, Yuan GC, Orkin SH. Combinatorial assembly of developmental stage-specific enhancers controls gene expression programs during human erythropoiesis. Dev Cell. 2012 Oct 16;23(4):796-811.

Ye T, Krebs AR, Choukrallah MA, Keime C, Plewniak F, Davidson I, Tora L. seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res.;39(6):e35.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-seq data. Bioinformatics. 2009 Aug 1;25(15):1952-8.

Zhang, Z. et al. PseudoPipe: an automated pseudogene identification pipeline. Bioinformatics (Oxford, England)22, 1437–9 (2006).

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C,Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-seq (MACS). Genome Biol. 2008;9(9):R137.

Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics. 2010 May 11;11:237.